

# Refitted Dimension-reduced tests for Cox's models based on regularization\*

Bo Tang

Department of Mathematics and Statistics  
University of North Carolina at Charlotte  
Email: btang2@charlotte.edu

## Abstract

We investigate hypothesis testing in Cox proportional hazards models under non-polynomial (NP) dimensionality, where the ambient dimension can exceed any fixed polynomial in the sample size. We introduce a DR-PLR test, together with a refitted version designed to temper the tendency to over-select variables. The method first fits a penalized partial likelihood to define a data-adaptive, low-dimensional working alternative space, and then conducts inference within that space using a PLR statistic. Under standard regularity conditions and appropriately calibrated regularization, the procedure achieves oracle-like performance and is insensitive to moderate changes in the tuning level. To curb spurious inclusions that can arise when  $p \gg n$ , we refit on the selected subset prior to testing. Theoretically, we derive Bahadur-type first-order expansions for the penalized estimators and characterize the asymptotic null distribution of the DR-PLR statistic. Monte Carlo experiments show consistent finite-sample gains over contemporary competitors, and a case study on non-Hodgkin's lymphoma survival demonstrates the workflow.

**Key words and phrases:** non-polynomial dimensionality, dimension-reduced partial likelihood ratio, penalized partial likelihood, refitting safeguard, Cox proportional hazards

## 1 Introduction

We observe i.i.d. samples  $\mathbf{v}_1, \dots, \mathbf{v}_n$  from the model  $f(\mathbf{v}; \boldsymbol{\theta})$  with parameter  $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ . Under the null,  $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ . We examine the composite null  $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$  against  $H_a : \boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$ . Let  $\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(\mathbf{v}_i; \boldsymbol{\theta})$ . The generalized likelihood-ratio statistic is

$$T_n = 2 \left\{ \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell_n(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \ell_n(\boldsymbol{\theta}) \right\}.$$

In the presence of nuisance components, a widely used formulation is

$$H_{0,1} : \boldsymbol{\gamma} = \mathbf{r} \quad \text{versus} \quad H_{a,1} : \boldsymbol{\gamma} \neq \mathbf{r}, \tag{1}$$

---

\*Correspondence should be addressed to

where  $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ ,  $\boldsymbol{\beta} \in \mathbb{R}^{p_n}$  are nuisance parameters and  $\boldsymbol{\gamma} \in \mathbb{R}^{q_n}$  are parameters of interest. Under the usual regularity conditions and with fixed  $q_n$ , the LR statistic satisfies  $T_n \Rightarrow \chi_{q_n}^2$  under  $H_{0,1}$ , and notably the limit does not involve  $\boldsymbol{\beta}$  (Wilks, 1938). Although classical arguments justify the local optimality of the LR via efficient MLEs, this relies on correct specification and accurate estimation; if the working model is unnecessarily large relative to the truth, weak nuisance effects can dilute power, and randomness induced by selection further complicates inference.

The analysis is conducted under the Cox PH model, using the partial likelihood for inference. In many modern studies,  $p_n$  can greatly exceed  $n$ . Regularization provides a practical screening device, but in ultrahigh dimensions, penalized fits may admit noise variables with non-negligible probability, enlarging the working model and weakening likelihood-type tests. We therefore treat regularization as a means to shape a data-driven target set on which the test can concentrate its power, rather than as the final goal.

We introduce a dimension-reduced partial likelihood ratio procedure (DR-PLR) for Cox models operating under non-polynomial (NP) dimensionality. The procedure begins with a penalized partial-likelihood fit that selects data-supported coordinates and thereby defines a low-dimensional working alternative; inference is then performed in that reduced space using a PLR statistic for testing (1). Empirically, once the selected set stabilizes across a range of regularization levels, the resulting PLR statistic varies little with moderate tuning changes, which helps concentrate power along data-supported directions.

We also cover the linear family:

$$H_{0,2} : \mathbf{A}\boldsymbol{\gamma} - \mathbf{A}\mathbf{r} = \mathbf{0} \quad \text{vs} \quad H_{a,2} : \mathbf{A}\boldsymbol{\gamma} - \mathbf{A}\mathbf{r} \neq \mathbf{0}, \quad (2)$$

with  $\mathbf{A}$  of full row rank  $m_n$  and a known target  $\mathbf{r}$ ; this includes testing membership of  $\boldsymbol{\gamma}$  in a  $(q_n - m_n)$ -dimensional subspace and reduces to (1) when  $m_n = q_n$ . Equivalently, because  $\mathbf{A}$  has full row rank, its null space has dimension  $q_n - m_n$ . By the SVD, there exists  $\mathbf{C} \in \mathbb{R}^{(q_n - m_n) \times q_n}$  with orthonormal rows spanning  $\ker(\mathbf{A})$ , i.e.,  $\mathbf{C}\mathbf{C}^\top = \mathbf{I}_{q_n - m_n}$  and  $\mathbf{A}\mathbf{C}^\top = \mathbf{0}$ , where  $\mathbf{I}_{q_n - m_n}$  is the identity matrix of size  $(q_n - m_n) \times (q_n - m_n)$ . We can write the null hypothesis as

$$H_{0,2} : \boldsymbol{\gamma} = \mathbf{r} + \mathbf{C}^\top \boldsymbol{\alpha},$$

For an unknown vector  $\boldsymbol{\alpha} \in \mathbb{R}^{q_n - m_n}$  (cf. Jiang and Jiang, 2011). The dimension  $p_n$  is allowed to grow faster than any fixed polynomial in  $n$  (e.g.,  $p_n \leq \exp\{n^\kappa\}$  with  $\kappa \in (0, 1)$ ), and  $(m_n, q_n)$  may also diverge at rates specified later. In special cases, our setting links to earlier partial-likelihood-based tests (see, e.g., Shi et al., 2019), which use partial penalized LR tests, and reduces to Fan and Peng (2004) for  $m_n$  held constant and  $p_n + q_n = o(n^{1/5})$ .

A convenient feature of PLR is that inference proceeds without modeling the baseline hazard; robustified versions are available, and DR-PLR keeps this insensitivity while shrinking the alternative through a screening step. To curb accidental inclusions from the screening stage, we add a refitting step that re-estimates the model on the selected coordinates before testing; analogous refitted Rao score and Wald versions can be formulated in the same spirit. These refitted procedures preserve the large-sample behavior of their non-refitted counterparts across broad practical tuning ranges while avoiding spurious effects from over-selection.

A key difficulty is the dependence of DR-PLR on a random post-screening model size. We handle this by deriving Bahadur-type first-order expansions for MPLEs in dimension-reduced models whose sizes are random. These expansions lead to a quadratic approximation: the PLR statistic is asymptotically equivalent to the squared norm of a suitably standardized difference of MPLEs under the reduced null and alternative, which delivers the asymptotic null law. We further obtain an explicit upper bound on the post-screening model size under NP growth, clarifying how screening interacts with inference.

Computationally, the pipeline is light: standard penalized Cox solvers on a tuning grid for screening, followed by a low-dimensional PLR evaluation. Monte Carlo studies show accurate size and competitive power over wide tuning ranges in NP regimes. The remainder of the article develops the DR-PLR construction and asymptotic theory for Cox models (including a limit on the post-screening model size); introduces refitted Rao and Wald versions together with their asymptotic limits; and reports simulation results and an application to non-Hodgkin's lymphoma survival. All technical conditions and proofs appear in the Appendix.

## 2 DR-PLR Testing for Cox Proportional Hazards Model

### 2.1 DR-PLR Test Statistic

To make construction concrete, we start with the DR-PLR procedure under Cox's proportional hazards model (Cox, 1972), which postulates the conditional hazard of a survival time  $T$  conditional on  $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})$  as

$$\lambda(t \mid \mathbf{x}, \mathbf{z}) = \lambda_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{x} + \boldsymbol{\gamma}^\top \mathbf{z}), \quad (3)$$

where  $\lambda_0(t)$  denotes the hazard and handle it as a nuisance term. Let  $\{(\mathbf{x}_i, \mathbf{z}_i, S_i, \delta_i)\}_{i=1}^n$  be an iid, with  $S_i$  is the observed time which is the minimum of censoring times  $C_i$  and survival time  $T_i$  and the event indicator  $\delta_i = \mathbf{1}\{T_i \leq C_i\}$ . We assume that  $\{C_i\}$  and  $\{T_i\}$  are independent conditional on the covariates  $(\mathbf{X}_i, \mathbf{Z}_i)$ , i.e.,  $C_i \perp T_i \mid (\mathbf{X}_i, \mathbf{Z}_i)$ . Arrange the observed event times

in increasing order  $t_1 < \dots < t_N$  and let  $\mathbf{X}_{(j)}$  be the covariate vector recorded at time  $t_j$ , where  $(j)$  be the index of the unit that fails at time  $t_j$ . Let  $\mathcal{R}_j = \{i : S_i \geq t_j\}$  be the risk set at  $t_j$ . The log partial likelihood is

$$Q_n(\boldsymbol{\theta}) = \sum_{j=1}^N j = 1^N \left\{ \boldsymbol{\theta}^\top \mathbf{w}_{(j)} - \log \left( \sum_{i \in \mathcal{R}_j} \exp(\boldsymbol{\theta}^\top \mathbf{w}_i) \right) \right\},$$

where  $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$  and  $\mathbf{w}_i = (\mathbf{x}_i^\top, \mathbf{z}_i^\top)^\top$ . The maximum partial likelihood estimator (MPLE) solves

$$\frac{\partial Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{j=1}^N \left\{ \mathbf{w}_{(j)} - \frac{\sum_{i \in \mathcal{R}_j} \mathbf{w}_i \exp(\boldsymbol{\theta}^\top \mathbf{w}_i)}{\sum_{i \in \mathcal{R}_j} \exp(\boldsymbol{\theta}^\top \mathbf{w}_i)} \right\} = 0. \quad (4)$$

It is convenient to adopt the following notation. Assume  $N_i(t) = \delta_i \mathbf{1}\{t \geq S_i\}$ ,  $Y_i(t) = \mathbf{1}\{t \leq S_i\}$ , and  $\bar{N}(t) = \sum_{i=1}^n \delta_i \mathbf{1}\{t \geq S_i\}$ . Assume the risk-set indicator vector  $\mathbf{Y}(t) = (Y_1(t), \dots, Y_n(t))^\top$  be LCRL and  $\forall t \in [0, \tau] : \Pr\{Y(t) = 1\} > 0$ . Then

$$Q_n(\boldsymbol{\theta}) = \sum_{i=1}^n \int_0^\tau \left\{ \boldsymbol{\theta}^\top \mathbf{w}_i(t) - \log S_n^{(0)}(\boldsymbol{\theta}, t) \right\} dN_i(t),$$

with

$$S_n^{(\ell)}(\boldsymbol{\theta}, t) = \sum_{i=1}^n Y_i(t) \{\mathbf{w}_i(t)\}^{\otimes \ell} \exp\{\boldsymbol{\theta}^\top \mathbf{w}_i(t)\}, \quad \ell = 0, 1, 2,$$

and  $\otimes$  denoting outer product. Define

$$\mathbf{E}_n(\boldsymbol{\theta}, t) = \frac{S_n^{(1)}(\boldsymbol{\theta}, t)}{S_n^{(0)}(\boldsymbol{\theta}, t)}, \quad \mathbf{V}_n(\boldsymbol{\theta}, t) = \frac{S_n^{(2)}(\boldsymbol{\theta}, t)}{S_n^{(0)}(\boldsymbol{\theta}, t)} - \{\mathbf{E}_n(\boldsymbol{\theta}, t)\}^{\otimes 2}.$$

Writing  $\mathbf{w}_i(t) = (\mathbf{x}_i(t)^\top, \mathbf{z}_i(t)^\top)^\top$  and  $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ , we also use the blockwise means

$$\mathbf{E}_n^{(1)}(\boldsymbol{\theta}, t) = \frac{S_{n1}^{(1)}(\boldsymbol{\theta}, t)}{S_n^{(0)}(\boldsymbol{\theta}, t)}, \quad \mathbf{E}_n^{(2)}(\boldsymbol{\theta}, t) = \frac{S_{n2}^{(1)}(\boldsymbol{\theta}, t)}{S_n^{(0)}(\boldsymbol{\theta}, t)},$$

where  $S_{n1}^{(1)}(\boldsymbol{\theta}, t) = \sum_{i=1}^n Y_i(t) \mathbf{x}_i(t) \exp\{\boldsymbol{\theta}^\top \mathbf{w}_i(t)\}$  and  $S_{n2}^{(1)}(\boldsymbol{\theta}, t) = \sum_{i=1}^n Y_i(t) \mathbf{z}_i(t) \exp\{\boldsymbol{\theta}^\top \mathbf{w}_i(t)\}$ .

The score and observed information admit the familiar forms

$$\frac{\partial Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \int_0^\tau \left\{ \mathbf{w}_i(t) - \mathbf{E}_n(\boldsymbol{\theta}, t) \right\} dN_i(t), \quad (5)$$

$$\frac{\partial^2 Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = - \sum_{i=1}^n \int_0^\tau \mathbf{V}_n(\boldsymbol{\theta}, t) dN_i(t). \quad (6)$$

For variable screening, we maximize a penalized partial log-likelihood (e.g., Bradic et al., 2011):

$$\mathcal{C}(\boldsymbol{\theta}, \tau) = Q_n(\boldsymbol{\theta}) - n \sum_{j=1}^p p_{\lambda_n}(|\theta_j|), \quad (7)$$

and choose any maximizer  $\hat{\boldsymbol{\theta}}$  of  $\mathcal{C}(\boldsymbol{\theta}, \tau)$  on the compact set  $\Omega_p \subset \mathbb{R}^p$  that contains the true value  $\boldsymbol{\theta}^*$ ; we refer to  $\hat{\boldsymbol{\theta}}$  as the sparse estimator. Let the individual intensity be  $\lambda_i(t, \boldsymbol{\theta}) = \lambda_0(t)Y_i(t) \exp\{\boldsymbol{\theta}^\top \mathbf{w}_i(t)\}$  and  $\Lambda_i(t) = \int_0^t \lambda_i(u, \boldsymbol{\theta}^*) du$ , where  $\boldsymbol{\theta}^*$  is the true value of  $\boldsymbol{\theta}$  with  $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*\top}, \boldsymbol{\gamma}^{*\top})^\top$ . Define  $M_i(t) = N_i(t) - \Lambda_i(t)$ . Under information set  $\mathcal{F}_t = \sigma\{N_i(u), \mathbf{w}_i(u^+), Y_i(u^+) : 0 \leq u \leq t, 1 \leq i \leq n\}$ ,  $\{M_i\}_{i=1}^n$  are locally square-integrable martingales that are orthogonal which means that  $\langle M_i, M_j \rangle_t = 0$  for all  $i \neq j$ . These tools, together with arguments akin to Fan and Li (2002) and Bradic et al. (2011), underlie the asymptotic theory developed later; the same construction extends to additive hazards (Jiang and Zhou, 2007) and transformation models (Doksum, 1987). This is also among the topics of this proposal.

For the testing task (1), the DR-PLR statistic is assembled via two steps:

- (i) *Screen the alternative.* Obtain the partially penalized estimator

$$\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}^\top, \tilde{\boldsymbol{\gamma}}^\top)^\top = \arg \max_{\boldsymbol{\theta}} \left\{ Q_n(\boldsymbol{\theta}) - n \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|) \right\}, \quad (8)$$

where only the nuisance block  $\boldsymbol{\beta}$  is penalized. Partition  $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^\top, \tilde{\boldsymbol{\beta}}_2^\top)^\top$  into its non-vanishing and vanishing coordinates, and split  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$  accordingly. Writing  $Q_n(\boldsymbol{\theta}) \equiv Q_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma})$ , we set  $\boldsymbol{\beta}_2 = \mathbf{0}$  and hence we work with the reduced partial likelihood  $Q_n(\boldsymbol{\beta}_1, \mathbf{0}, \boldsymbol{\gamma})$ .

- (ii) *Form the PLR on the reduced model.* Let

$$\hat{\boldsymbol{\theta}}_0 = (\hat{\boldsymbol{\beta}}_{1,0}^\top, \mathbf{0}^\top, \mathbf{r}^\top)^\top, \quad \hat{\boldsymbol{\theta}}_a = (\hat{\boldsymbol{\beta}}_{1,a}^\top, \mathbf{0}^\top, \hat{\boldsymbol{\gamma}}_a^\top)^\top$$

be the MPLEs under  $H_{0,1}$  and  $H_{a,1}$  for the reduced model. Define

$$T_{n,1} = 2\{Q_n(\hat{\boldsymbol{\theta}}_a) - Q_n(\hat{\boldsymbol{\theta}}_0)\}. \quad (9)$$

Large values of  $T_{n,1}$  support rejection of  $H_{0,1}$ .

Two remarks are worth noting. First, Step (i) penalizes only the nuisance coefficients, so no beta-min requirement is imposed on the components fixed by the null; weak nuisance signals remain identifiable and underfitting is reduced. Second, unless no reduction occurs, the MPLEs computed in Step (ii) are not the full-model MPLEs; consequently,  $T_{n,1}$  is not the classical LR statistic for the unreduced model (in particular when some nuisance parameters are zero). If Step (i) selects no unimportant covariates, then  $T_{n,1}$  coincides with  $T_n$ . To avoid missing relevant effects, we require the penalized estimator in (8) to possess the sure-screening property (Fan and Lv, 2008), which is readily achieved in practice (see also Fan, Ning and Hao, 2012).

The framework extends beyond exact likelihoods: the statistic is estimator-driven, so we continue to refer to the resulting procedures as DR-PLR tests. Concretely, in (8) one may replace  $Q_n(\boldsymbol{\theta})$  with the negative of a suitable loss, thereby accommodating quasi-/pseudo-likelihood criteria (see Fan et al., 2001; Fan and Jiang, 2007). The penalty  $p_{\lambda_n}(\cdot)$  is governed by a tuning level  $\lambda_n$ , which can be selected via information criteria (e.g., BIC). Examples include ridge/Tikhonov, the nonnegative garrote (Yuan and Lin, 2007),  $\ell_0$ , lasso (Tibshirani, 1996), and folded-concave penalties—SCAD (Fan and Li, 2001). For testing purposes, penalties with strong selection behavior—such as adaptive lasso (Zou, 2006) and SCAD/MCP (Kim et al., 2008; Zou and Yuan, 2008)—are particularly appealing.

In Step (i), we apply shrinkage only to the nuisance coordinates, which prunes the alternative while leaving the parameters under test untouched. This design serves two aims: it concentrates power by reducing the dimensionality on the alternative side, and it preserves nominal size under  $H_0$  by avoiding shrinkage on the target coordinates. In practice, one fits a Cox proportional hazards model with penalties on nuisance indices only, carries the selected nuisance set to Step (ii), and keeps the parameters of interest in the reduced model so that the subsequent DR-PLR statistic is well defined.

Intuitively, by screening out negligible nuisance coefficients on the alternative side, the procedure effectively emulates an oracle that knows the zero components a priori; when the posited model is correct up to a sparse nuisance block—a typical high-dimensional regime—the reduction step yields marked power gains over the unreduced likelihood-ratio (LR) test while keeping type-I error at the nominal level.

Actually, DR-PLR chooses a single  $\lambda_n$  in Step (i). Once the screened set stabilizes, the Step (ii) statistic varies little over a broad range of  $\lambda_n$ , yielding stable size and power (see §4). Moreover, the  $\lambda_{\text{opt}}$  that performs well for estimation typically also works well for testing, simplifying calibration—an advantage not commonly enjoyed by alternative regularization-based tests.

For the general linear hypothesis in (2),

$$H_{0,2} : \mathbf{A}\boldsymbol{\gamma} - \mathbf{A}\mathbf{r} = \mathbf{0} \quad \text{versus} \quad H_{a,2} : \mathbf{A}\boldsymbol{\gamma} - \mathbf{A}\mathbf{r} \neq \mathbf{0},$$

the estimator under  $H_{a,2}$  remains  $\hat{\boldsymbol{\theta}}_a = (\hat{\boldsymbol{\beta}}_{1,a}^\top, \mathbf{0}^\top, \hat{\boldsymbol{\gamma}}_a^\top)^\top$  with  $(\hat{\boldsymbol{\beta}}_{1,a}, \hat{\boldsymbol{\gamma}}_a) = \arg \max Q_n(\boldsymbol{\beta}_1, 0, \boldsymbol{\gamma})$ . Under  $H_{0,2}$ ,

$$\hat{\boldsymbol{\theta}}_0 = (\hat{\boldsymbol{\beta}}_{1,0}^\top, \mathbf{0}^\top, \hat{\boldsymbol{\gamma}}_0^\top) = \arg \max_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \{ Q_n(\boldsymbol{\beta}_1, 0, \boldsymbol{\gamma}) : \mathbf{A}\boldsymbol{\gamma} - \mathbf{A}\mathbf{r} = \mathbf{0} \}.$$

We maximize  $Q_n(\boldsymbol{\beta}_1, \mathbf{0}, \boldsymbol{\gamma})$  with  $\boldsymbol{\gamma} = \mathbf{r} + \mathbf{C}^\top \boldsymbol{\alpha}$ , and obtain the MPLE

$$\hat{\boldsymbol{\theta}}_0 = (\hat{\boldsymbol{\beta}}_{1,0}^\top, \mathbf{0}^\top, (\mathbf{r} + \mathbf{C}^\top \hat{\boldsymbol{\alpha}})^\top)^\top. \quad (10)$$

The corresponding statistic is

$$T_{n,2} = 2\{Q_n(\hat{\boldsymbol{\theta}}_a) - Q_n(\hat{\boldsymbol{\theta}}_0)\}, \quad (11)$$

which reduces to  $T_{n,1}$  when  $\mathbf{A}$  is invertible.

## 2.2 Large-Sample Theory for the DR-PLR Procedure

Now, we develop the large-sample distributional properties of DR-PLR. A key prerequisite is to control how large the post-screening working alternative can be; this “effective dimension” drives both validity and power of the statistic.

**Notation.** For clarity, we fix notation. Let  $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_{p_n}^*)^\top$  be the true parameter vector of  $\boldsymbol{\beta}$ ,  $\boldsymbol{\gamma}^* = (\gamma_1^*, \dots, \gamma_{q_n}^*)^\top$  be the true parameter vector of  $\boldsymbol{\gamma}$ ; then we have  $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*\top}, \boldsymbol{\gamma}^{*\top})^\top \in \mathbb{R}^{p_n+q_n}$ . Let  $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_n)^\top \in \mathbb{R}^{n \times (p_n+q_n)}$  be the covariate matrix, where  $\mathbf{w}_i = (\mathbf{x}_i^\top, \mathbf{z}_i^\top)^\top$  stacks the nuisance and target covariates. Let  $\mathbf{D} \subset \{1, \dots, p_n + q_n\}$  index the coordinates of  $\boldsymbol{\theta}$  corresponding to  $\boldsymbol{\gamma}$ , so that  $\boldsymbol{\theta}_{\mathbf{D}}^* = \boldsymbol{\gamma}^*$ . Define the true support of the nuisance block by  $\mathbf{S} = \{j \in \{1, \dots, p_n\} : \beta_j^* \neq 0\}$  and the selected set by  $\hat{\mathbf{S}} = \{j \leq p_n : \tilde{\beta}_j \neq 0\}$ ; let  $\hat{\mathbf{N}} = \hat{\mathbf{S}} \setminus \mathbf{S}$  denote false inclusions. Write  $s_n = |\mathbf{S}|$  and define the minimal signal level  $d_n = \frac{1}{2} \min_{j \in \mathbf{S}} |\beta_j^*|$  (with the convention  $d_n = \infty$  if  $\mathbf{S} = \emptyset$ ). For any index set  $\mathbf{N}$ ,  $|\mathbf{N}|$  denotes its cardinality. Let  $Q$  denote the working criterion (e.g., the Cox partial log-likelihood), and  $g$  the link if applicable (for Cox,  $g(t) \equiv t$ ). Define

$$v_1(t, y) = \partial Q(g^{-1}(t), y) / \partial t, \quad v_2(t, y) = \partial^2 Q(g^{-1}(t), y) / \partial t^2.$$

and collect  $\mathbf{v}_1(\boldsymbol{\theta}) = (v_1(\mathbf{w}_1^\top \boldsymbol{\theta}, Y_1), \dots, v_1(\mathbf{w}_n^\top \boldsymbol{\theta}, Y_n))^\top$ . For a vector  $\mathbf{a}$ , let  $\|\mathbf{a}\|_\infty = \max_j |a_j|$  and  $\|\mathbf{a}\|_0 = \#\{j : a_j \neq 0\}$ . If  $\boldsymbol{\xi}$  is a vector and  $\mathbf{J}$  an index set, then  $\boldsymbol{\xi}_{\mathbf{J}}$  denotes the subvector of  $\boldsymbol{\xi}$  with indices in  $\mathbf{J}$ .

For sequences  $\{\xi_n\}$  and  $\{\eta_n\}$ , denote  $\xi_n \stackrel{a}{\sim} \eta_n$  whenever  $\sup_x |P(\xi_n \leq x) - P(\eta_n \leq x)| \rightarrow 0$  as  $n \rightarrow \infty$ . We write  $\chi_{m_n}^2(\eta_n^2)$  for the chi-square law on  $m_n$  degrees of freedom with noncentrality parameter  $\eta_n^2$ ; the central case ( $\eta_n^2 = 0$ ) is written  $\chi_{m_n}^2$ .

**Theorem 1.** Under conditions  $(A_1)$ – $(A_5)$  and  $\frac{s_n + q_n}{n^{1/2}} \rightarrow 0$  as  $n \rightarrow \infty$ , the number of false inclusions satisfies

$$|\hat{\mathbf{N}}| = O_p(s_n + q_n).$$

If, in addition,  $(A_6)$  holds, no spurious variables are retained w.h.p., i.e.,  $|\hat{\mathbf{N}}| = 0$ .

We next state Wilks-type null limits for the ordinary partial-likelihood ratio statistic  $T_n$  and for its dimension-reduced counterpart  $T_2$  aimed at testing  $H_{0,2} : \mathbf{A}(\boldsymbol{\gamma} - \mathbf{t}) = \mathbf{0}$ .

**Theorem 2.** Under conditions  $(A_1)$ – $(A_6)$ ,  $(B_1)$ ,  $(B_2)$ , and  $(B_4)$ , there exist constants  $C_1, C_2 \in (0, \infty)$ , independent of  $n$ , such that  $\|\mathbf{A}\mathbf{A}^\top\|_{\text{op}} \leq C_1$  and  $\|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_{\text{op}} \leq C_2$ . Under the null  $H_{0,2}$ , we have

(i) If  $\frac{p_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), then  $\lambda_n$ ,  $T_{n,2}$ , and  $\chi_{m_n}^2$  are asymptotically identically distributed, which is denoted by  $T_n \stackrel{a}{\sim} \chi_{m_n}^2$ .

(ii) If  $\frac{s_n + q_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), then  $T_{n,2} \stackrel{a}{\sim} \chi_{m_n}^2$ .

**Remark 1.** Under non-polynomial growth, DR-PLR preserves the Wilks phenomenon provided the effective post-screening dimension is  $o(n^{1/3})$ .

For power analysis of the DR-PLR test, we consider Pitman local alternatives contiguous to  $H_{0,2}$ ,

$$H_{a,2}^{(n)} : \boldsymbol{\gamma} = \mathbf{r} + \mathbf{C}^\top \boldsymbol{\alpha} + \tilde{\boldsymbol{\Delta}}_n, \quad \|\tilde{\boldsymbol{\Delta}}_n\|_2 \rightarrow 0,$$

and assume there is no  $\boldsymbol{\alpha}^* \in \mathbb{R}^{q_n - m_n}$  with  $\tilde{\boldsymbol{\Delta}}_n = \mathbf{C}^\top \boldsymbol{\alpha}^*$  (otherwise  $H_{a,2}^{(n)}$  coincides with  $H_{0,2}$  and no test has power). Assume  $\text{Null}(\mathbf{A}) = \text{col}(\mathbf{C}^\top)$ —equivalently,  $\text{rank}(\mathbf{A}) = m_n$ ,  $\text{rank}(\mathbf{C}) = q_n - m_n$ , and  $\mathbf{A}\mathbf{C}^\top = \mathbf{0}$ . With  $\boldsymbol{\Delta}_n := \mathbf{A}\tilde{\boldsymbol{\Delta}}_n$ , this is equivalently written as

$$\mathbf{A}\boldsymbol{\gamma} - \mathbf{A}\mathbf{r} = \boldsymbol{\Delta}_n, \quad \|\boldsymbol{\Delta}_n\|_2 \neq 0 \text{ for all sufficiently large } n.$$

To handle local alternatives, we first set notation and then characterize the large-sample laws of the LR and DR-PLR tests. Define the partial-likelihood information matrix

$$\mathbf{I}(\boldsymbol{\theta}) = E\{-N^{-1}\partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top\} = E\left\{-N^{-1}\begin{pmatrix} \partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top & \partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^\top \\ \partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}^\top & \partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top \end{pmatrix}\right\}$$

and suppose  $\mathbf{I}(\boldsymbol{\theta}^*)$  is invertible. In accordance with the split  $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ , we block-decompose  $\mathbf{I}(\boldsymbol{\theta})$  and its inverse  $\mathbf{I}^{-1}(\boldsymbol{\theta})$  accordingly as

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{I}_{11}(\boldsymbol{\theta}) & \mathbf{I}_{12}(\boldsymbol{\theta}) \\ \mathbf{I}_{21}(\boldsymbol{\theta}) & \mathbf{I}_{22}(\boldsymbol{\theta}) \end{pmatrix} \quad \text{and} \quad \mathbf{I}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{I}^{11}(\boldsymbol{\theta}) & \mathbf{I}^{12}(\boldsymbol{\theta}) \\ \mathbf{I}^{21}(\boldsymbol{\theta}) & \mathbf{I}^{22}(\boldsymbol{\theta}) \end{pmatrix}$$

where  $\mathbf{I}(\boldsymbol{\theta})$  is  $p_n \times p_n$  and  $\mathbf{I}^{-1}(\boldsymbol{\theta})$  is  $q_n \times q_n$ . Assume  $j_1, \dots, j_{s_n}$  are the indices in  $\mathbf{S}$  arranged in increasing order, and let  $\mathbf{e}$  be an  $s_n \times p_n$  matrix with the  $j_l$ -th column equal to the  $l$ -th standard basis vector ( $l = 1, \dots, s_n$ ) and the remaining columns equal to zero in the Euclidean

space  $R^{s_n}$ . Let  $I_r(\boldsymbol{\theta})$  denote the reduced information matrix obtained from  $I(\boldsymbol{\theta})$  with indices lying in  $\mathcal{M}_0$ . That is,

$$\mathbf{I}_o(\boldsymbol{\theta}) = E\left\{-N^{-1}\begin{pmatrix} \partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\beta}_s\partial\boldsymbol{\beta}_s^\top & \partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\beta}_s\partial\boldsymbol{\gamma}^\top \\ \partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}_s^\top & \partial^2 Q_n(\boldsymbol{\theta})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top \end{pmatrix}\right\}.$$

with  $\mathbf{I}_o^{-1}(\boldsymbol{\theta})$  partitioned in the same way.

$$\mathbf{I}_o(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{I}_{o,11}(\boldsymbol{\theta}) & \mathbf{I}_{o,12}(\boldsymbol{\theta}) \\ \mathbf{I}_{o,21}(\boldsymbol{\theta}) & \mathbf{I}_{o,22}(\boldsymbol{\theta}) \end{pmatrix} \quad \text{and} \quad \mathbf{I}_o^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{I}_o^{11}(\boldsymbol{\theta}) & \mathbf{I}_o^{12}(\boldsymbol{\theta}) \\ \mathbf{I}_o^{21}(\boldsymbol{\theta}) & \mathbf{I}_o^{22}(\boldsymbol{\theta}) \end{pmatrix}$$

The associated Schur complement satisfies

$$\{\mathbf{I}_o^{22}(\boldsymbol{\theta})\}^{-1} = \mathbf{I}_{o,22}(\boldsymbol{\theta}) - \mathbf{I}_{o,21}(\boldsymbol{\theta})\{\mathbf{I}_{o,11}(\boldsymbol{\theta})\}^{-1}\mathbf{I}_{o,12}(\boldsymbol{\theta}).$$

**Theorem 3.** (*Local-alternative limits*) Under the assumptions of Theorem 2 and with  $\|\boldsymbol{\Delta}_n\|_2 = O(\sqrt{m_n/n})$ :

(i) If  $\frac{p_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), then

$$T_n \stackrel{a}{\sim} \chi_{m_n}^2(\eta_n^2), \quad \eta_n^2 = n \boldsymbol{\Delta}_n^\top \{\mathbf{A} \mathbf{I}^{22}(\boldsymbol{\theta}^*) \mathbf{A}^\top\}^{-1} \boldsymbol{\Delta}_n.$$

(ii) If  $\frac{s_n+q_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), then

$$T_{n,2} \stackrel{a}{\sim} \chi_{m_n}^2(\eta_{n,o}^2), \quad \eta_{n,o}^2 = n \boldsymbol{\Delta}_n^\top \{\mathbf{A} \mathbf{I}_o^{22}(\boldsymbol{\theta}^*) \mathbf{A}^\top\}^{-1} \boldsymbol{\Delta}_n.$$

**Corollary 1.** If  $\mathbf{I}(\boldsymbol{\theta}^*)$  is invertible, then  $\eta_n^2 \leq \eta_{n,o}^2$ , with equality iff there is no nuisance block or the cross-information between  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$  vanishes.

According to Corollary 1, whenever at least one nuisance coordinate is truly zero, the large-sample power of  $T_{n,2}$  strictly exceeds the large-sample power of  $T_n$ , except when there is no nuisance block or the cross-information between  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$  vanishes. The gain is maximized when all nuisance effects vanish. If no nuisance entry is zero, then  $\lambda_n$  and  $T_{n,1}$  are asymptotically equivalent. Moreover, if the zero nuisance coordinates were known *a priori* and one formed the LR test after removing them, its large-sample power would match that of DR-PLR; thus DR-PLR enjoys an oracle property. Consequently, when many nuisance parameters are inactive, DR-PLR effectively shrinks the alternative space and achieves higher power than LR.

In high-dimensional settings with many nuisance parameters—such as breast cancer studies with genome-wide expression—one tests the joint significance of a target gene group while adjusting for the remaining genes. When most nuisance effects are zero, DR-PLR typically attains substantially greater power than PLR-type procedures by reducing the effective alternative dimension.

### 3 Refitted tests

#### 3.1 Refitted DR-PLR test

In ultra-high-dimensional regression, achieving “sure screening” typically leads selectors to err on the side of inclusion; as documented by Fan, Ning and Hao (2012), noise covariates can be admitted with non-negligible probability. If the penalized fit in (8) carries such false positives into the screened set, a natural question is how this impacts our DR-PLR: does it distort size (inflate type-I error), erode power, or both? We investigate these issues via extensive Monte Carlo experiments under an extremely large ambient dimension. To blunt the influence of over-selection, we further introduce a *refitted* DR-PLR: after screening, we re-estimate on the retained coordinates and then form the test statistic, echoing the spirit of refitted cross-validation in Fan, Ning and Hao (2012). The refitted variant aims to purge spurious effects while preserving the tuning-insensitivity of the original DR-PLR.

For clarity of exposition, we revisit the testing problem  $H_{0,2}$ . We assume the sample is randomly split into two subsamples  $\{(\mathbf{w}_i^{(1)}, S_i^{(1)}, \delta_i^{(1)})\}_{i=1}^N$ , and  $\{(\mathbf{w}_i^{(2)}, S_i^{(2)}, \delta_i^{(2)})\}_{i=1}^N$ , where  $n = 2N$  and  $\mathbf{w}_i^{(k)} = (\{\mathbf{x}_i^{(k)}\}^\top, \{\mathbf{z}_i^{(k)}\}^\top)^\top$  for  $k = 1, 2$ . For the  $k$ th subsample, arrange the observed event times in increasing order  $t_1^{(k)} < \dots < t_N^{(k)}$  and let  $(j)$  be the index of the unit that fails at time  $t_j^{(k)}$ . Define the risk set by  $\mathcal{R}_j^{(k)} = \{i \in \{1, \dots, N\} : S_i^{(k)} \geq t_j^{(k)}\}$ ,  $j = 1, \dots, N$ . For dimension reduction, we fit model (3) to the  $k$ th subsample under  $H_{a,2}$  by

$$L_{n,\lambda}(\boldsymbol{\theta}; \mathbf{w}_1, \dots, \mathbf{w}_n) = \tilde{\boldsymbol{\theta}}^{(k)} = \max_{\boldsymbol{\theta}} \left\{ Q_n^{(k)}(\boldsymbol{\theta}) - N \sum_{j=1}^{p_n} p_{\lambda_n}(|\theta_j|) \right\},$$

where  $Q_n^{(k)}(\boldsymbol{\theta}) = \sum_{j=1}^N \{\boldsymbol{\theta}^\top \mathbf{w}_{(j)}^{(k)} - \log(\sum_{i \in \mathcal{R}_j^{(k)}} \exp(\boldsymbol{\theta}^\top \mathbf{w}_i^{(k)}))\}$ . Let  $\hat{\mathbf{S}}_k = \{i : \tilde{\theta}_i^{(k^*)} \neq 0, 1 \leq i \leq p_n\}$  and  $s_{nk} = |\hat{\mathbf{S}}_k|$ , where  $k^* = 2 - |k - 1|$  and  $k$  can be 1 or 2.

Let

$$\begin{aligned} Q_n^{(k)}(\boldsymbol{\theta}) &= \sum_{i=1}^N \int_0^\tau \{\boldsymbol{\theta}^\top \mathbf{w}_i^{(k)}(t) - \log(S_{n,k}^{(0)}(\boldsymbol{\theta}, t))\} dN_i^{(k)}(t), \\ S_{n,k}^{(\ell)}(\boldsymbol{\theta}, t) &= \sum_{i=1}^N Y_i^{(k)}(t) \{\mathbf{w}_i^{(k)}(t)\}^{\otimes \ell} \exp(\boldsymbol{\theta}^\top \mathbf{w}_i^{(k)}(t)) \end{aligned} \quad (12)$$

where  $\ell$  can be 0, 1, or 2, and  $\otimes$  means the dyadic product.

$$\mathbf{E}^{(k)}(\boldsymbol{\theta}, t) = S_{n,k}^{(1)}(\boldsymbol{\theta}, t)/S_{n,k}^{(0)}(\boldsymbol{\theta}, t) \text{ and } \mathbf{V}^{(k)}(\boldsymbol{\theta}, t) = S_{n,k}^{(2)}(\boldsymbol{\theta}, t)/S_{n,k}^{(0)}(\boldsymbol{\theta}, t) - \{\mathbf{E}^{(k)}(\boldsymbol{\theta}, t)\}^{\otimes 2}.$$

To reduce the impact of spurious inclusions on the DR-PLR statistic, we re-estimate the Cox model on the screened coordinates, enforcing  $H_{0,2}$  and  $H_{a,2}$  in turn. Concretely, for model (3) we compute the MPLEs under  $H_{0,2}$  and  $H_{a,2}$ , denoted by

$$\hat{\boldsymbol{\theta}}_{k,0} = \arg \max_{\boldsymbol{\theta}} \{Q_n^{(k)}(\boldsymbol{\theta}), \text{ s.t. } \boldsymbol{\beta}_{\hat{\mathbf{S}}_k^c} = \mathbf{0}, \mathbf{A}\boldsymbol{\gamma} - \mathbf{A}\mathbf{t} = \mathbf{0}\}$$

and

$$\hat{\boldsymbol{\theta}}_{k,a} = \arg \max_{\boldsymbol{\theta}} \{Q_n^{(k)}(\boldsymbol{\theta}), \text{ s.t. } \boldsymbol{\beta}_{\hat{\mathbf{S}}_k^c} = \mathbf{0}\},$$

respectively. Evaluating the partial log-likelihood at the null and alternative fits yields  $Q_n^{(k)}(\hat{\boldsymbol{\theta}}_{k,0})$  and  $Q_n^{(k)}(\hat{\boldsymbol{\theta}}_{k,a})$  for split  $k$ . In analogy with (11), we define the DR-PLR statistic on split  $k$  by

$$T_{n,\mathbf{L}}^{(k)} = 2\{Q_n^{(k,a)}(\hat{\boldsymbol{\theta}}_k) - Q_n^{(k)}(\hat{\boldsymbol{\theta}}_{k,0})\}.$$

A simple aggregate for the two splits is

$$T_{n,\mathbf{L}}^* = \frac{T_{n,\mathbf{L}}^{(1)} + T_{n,\mathbf{L}}^{(2)}}{2}.$$

However, even with ideal reduction, this aggregated statistic may still fall short of  $T_{n,2}$  in power. In fact, when  $\hat{\mathbf{S}}_k = \mathbf{S}$  we have  $T_{n,\mathbf{L}}^{(k)} \stackrel{a}{\sim} \chi_{m_n}^2$  and  $T_{n,\mathbf{L}}^* \stackrel{a}{\sim} \frac{1}{2}\chi_{2m_n}^2$ , so the effective degrees of freedom are doubled relative to  $T_{n,2}$  (see Ghosh, 1973), which tends to dilute power.

We now put forward a construction of the refitted DR-PLR statistic that aggregates information from the entire sample. The recipe hinges on a Bahadur-type linearization of the split-specific MPLEs  $\hat{\boldsymbol{\theta}}_k$  together with a second-order (quadratic) expansion of  $T_{n,\mathbf{L}}^{(k)}$ . For the ensuing derivations, we first set up some notation. Define

$$\mathbf{I}_k(\boldsymbol{\theta}) = -N^{-1}E \begin{pmatrix} \partial^2 Q_n^{(k)}(\boldsymbol{\theta}) / \partial \boldsymbol{\beta}_{\hat{\mathbf{S}}_k} \partial \boldsymbol{\beta}_{\hat{\mathbf{S}}_k}^\top & \partial^2 Q_n^{(k)}(\boldsymbol{\theta}) / \partial \boldsymbol{\beta}_{\hat{\mathbf{S}}_k} \partial \boldsymbol{\gamma}^\top \\ \partial^2 Q_n^{(k)}(\boldsymbol{\theta}) / \partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}_{\hat{\mathbf{S}}_k}^\top & \partial^2 Q_n^{(k)}(\boldsymbol{\theta}) / \partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\top \end{pmatrix},$$

and its empirical version

$$\hat{\mathbf{I}}_k(\boldsymbol{\theta}) = -N^{-1} \begin{pmatrix} \partial^2 Q_n^{(k)}(\boldsymbol{\theta}) / \partial \boldsymbol{\beta}_{\hat{\mathbf{S}}_k} \partial \boldsymbol{\beta}_{\hat{\mathbf{S}}_k}^\top & \partial^2 Q_n^{(k)}(\boldsymbol{\theta}) / \partial \boldsymbol{\beta}_{\hat{\mathbf{S}}_k} \partial \boldsymbol{\gamma}^\top \\ \partial^2 Q_n^{(k)}(\boldsymbol{\theta}) / \partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}_{\hat{\mathbf{S}}_k}^\top & \partial^2 Q_n^{(k)}(\boldsymbol{\theta}) / \partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\top \end{pmatrix}.$$

Partition  $\mathbf{I}_k^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{I}_k^{11}(\boldsymbol{\theta}) & \mathbf{I}_k^{12}(\boldsymbol{\theta}) \\ \mathbf{I}_k^{21}(\boldsymbol{\theta}) & \mathbf{I}_k^{22}(\boldsymbol{\theta}) \end{pmatrix}$  and  $\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \hat{\mathbf{I}}_k^{11}(\boldsymbol{\theta}) & \hat{\mathbf{I}}_k^{12}(\boldsymbol{\theta}) \\ \hat{\mathbf{I}}_k^{21}(\boldsymbol{\theta}) & \hat{\mathbf{I}}_k^{22}(\boldsymbol{\theta}) \end{pmatrix}$ , where both  $\mathbf{I}_k^{22}(\boldsymbol{\theta})$  and  $\hat{\mathbf{I}}_k^{22}(\boldsymbol{\theta})$  have dimension  $q_n \times q_n$ . Let  $\mathbb{W}_k$  be the split- $k$  model matrix, with nuisance and target submatrices denoted by  $\mathbb{X}_k$  and  $\mathbb{Z}_k$ , where  $\mathbb{X}_k = ((\mathbf{x}_{1,\hat{\mathbf{S}}_k}^{(k)})^\top, \dots, (\mathbf{x}_{N,\hat{\mathbf{S}}_k}^{(k)})^\top)^\top$  and  $\mathbb{Z}_k = ((\mathbf{z}_1^{(k)})^\top, \dots, (\mathbf{z}_N^{(k)})^\top)^\top$ . Let  $\mathbf{D} \subset \{1, \dots, p_n + q_n\}$  denote the index set selecting the target block in  $\boldsymbol{\theta}$ , so that  $\boldsymbol{\theta}_{\mathbf{D}}^* = \boldsymbol{\gamma}^*$ . Define  $\mathbf{T} = \mathbf{S} \cup \mathbf{D}$ , which collects the true support of the nuisance

part  $\beta$  together with all coordinates of  $\gamma^*$ . In addition, suppose  $\hat{\mathbf{T}}_k = \hat{\mathbf{S}}_k \cup \mathbf{D}$ ,  $\mathbf{T}_k = \mathbf{S}_k \cup \mathbf{D}$ ,  $\tilde{\mathbf{A}}_k = (\mathbf{0}_{m_n \times s_{nk}}, \mathbf{A})$ ,  $\mathbf{\Omega}_k(\theta^*) = \{\tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\theta^*) \tilde{\mathbf{A}}_k^\top\}^{-1}$ , and  $\theta_{\hat{\mathbf{T}}_k} = (\beta_{\hat{\mathbf{S}}_k}^\top, \gamma^\top)^\top$ , where  $k$  can be 1 or 2. Then  $|\hat{\mathbf{T}}_k| = \tau_{nk}$ . Obviously,

$$\mathbf{U}(\theta) \equiv \frac{\partial Q_n(\theta)}{\partial \theta} = \sum_{i=1}^n \int_0^\tau \{\mathbf{w}_i(t) - \mathbf{E}(\theta, t)\} dN_i(t) \quad (13)$$

and

$$\mathbf{H}(\theta) \equiv \frac{\partial^2 Q_n(\theta)}{\partial \theta \partial \theta^\top} = - \sum_{i=1}^n \int_0^\tau \mathbf{V}(\theta, t) dN_i(t). \quad (14)$$

Furthermore,

$$\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\theta) \equiv \frac{\partial Q_n^{(k)}(\theta)}{\partial \theta_{\hat{\mathbf{T}}_k}} = \sum_{i=1}^N \int_0^\tau \{\mathbf{w}_{i, \hat{\mathbf{T}}_k}^{(k)}(t) - \mathbf{E}_{\hat{\mathbf{T}}_k}^{(k)}(\theta, t)\} dN_i^{(k)}(t)$$

and

$$\mathbf{H}_{\hat{\mathbf{T}}_k, \hat{\mathbf{T}}_k}^{(k)}(\theta) \equiv \frac{\partial^2 Q_n^{(k)}(\theta)}{\partial \theta_{\hat{\mathbf{T}}_k} \partial \theta_{\hat{\mathbf{T}}_k}^\top} = - \sum_{i=1}^N \int_0^\tau \mathbf{V}_{\hat{\mathbf{T}}_k, \hat{\mathbf{T}}_k}^{(k)}(\theta, t) dN_i^{(k)}(t),$$

where  $\mathbf{V}_{\hat{\mathbf{T}}_k, \hat{\mathbf{T}}_k}^{(k)}(\theta, t)$  is the submatrix of  $\mathbf{V}^{(k)}(\theta, t)$  with row and column indexes in  $\hat{\mathbf{T}}_k$ . Finally, we obtain the following results.

**Theorem 4.** Under conditions  $(A_1)$ – $(A_5)$  and  $(B_1)$ . If  $\frac{s_n + q_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), and there exist constants  $C_1, C_2 \in (0, \infty)$ , independent of  $n$ , such that  $\|\mathbf{A}\mathbf{A}^\top\|_{\text{op}} \leq C_1$  and  $\|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_{\text{op}} \leq C_2$ , then under  $H_{a,2}^{(n)}$  with  $\|\Delta_n\|_2 = O(\sqrt{m_n/n})$ , where  $k$  can be 1 or 2, we have

(i) Sparse structure of the estimators:  $(\hat{\theta}_{k,a} - \theta^*)_{\hat{\mathbf{T}}_k^c} = \mathbf{0}$  and  $(\hat{\theta}_{k,0} - \theta^*)_{\hat{\mathbf{T}}_k^c} = \mathbf{0}$ ;

(ii) Asymptotic linear representation:

(1) Alternative estimator:  $\sqrt{N}(\hat{\theta}_{k,a} - \theta^*)_{\hat{\mathbf{T}}_k} = N^{-1/2} \mathbf{I}_k^{-1}(\theta^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\theta) + r_n$ , where  $\|r_n\|_2 = o_P(1)$ ;

(2) Null estimator:

$$\begin{aligned} \sqrt{N}(\hat{\theta}_{k,0} - \theta^*)_{\hat{\mathbf{T}}_k} &= N^{-1/2} \mathbf{I}_k^{-1}(\theta^*) \{\mathbb{I}_{\tau_{nk}} - \tilde{\mathbf{A}}_k^\top \mathbf{\Omega}_k(\theta^*) \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\theta^*)\} \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\theta) \\ &\quad - \mathbf{I}_k^{-1}(\theta^*) \tilde{\mathbf{A}}_k^\top \mathbf{\Omega}_k(\theta^*) \Delta_n + \mathbf{r}_n^*, \end{aligned}$$

with  $\|\mathbf{r}_n^*\|_2 \xrightarrow{P} 0$ , where  $\mathbb{I}_{q_n}$  denotes the identity matrix of order  $q_n$ ;

(iii) Second-order approximation:

$$2\{Q_n^{(k)}(\hat{\theta}_{k,a}) - Q_n^{(k)}(\hat{\theta}_{k,0})\} = \|\zeta_{n,k}\|_2^2 + o_P(m_n),$$

where  $\zeta_{n,k} = \sqrt{N}\{\mathbf{A}\hat{\mathbf{I}}_k^{22}(\hat{\boldsymbol{\theta}}_{k,a})\mathbf{A}^\top\}^{1/2}(\mathbf{A}\mathbf{A}^\top)^{-1}\tilde{\mathbf{A}}_k\hat{\mathbf{I}}_k(\hat{\boldsymbol{\theta}}_{k,a})(\hat{\boldsymbol{\theta}}_{k,a} - \hat{\boldsymbol{\theta}}_{k,0})_{\hat{\mathbf{T}}_k}$ . Moreover, under  $H_{0,2}$ , we have  $\|\zeta_{n,k} - \zeta_{n,k}^*\|_2 \xrightarrow{P} 0$ , where  $\zeta_{n,k}^* = \sqrt{N}\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)U_{n,\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta})$ . In addition, if  $(A_6)$  and  $(B_2)$  hold, then  $\zeta_{n,k}^* \stackrel{a}{\sim} N_{m_n}(\mathbf{0}, \phi\mathbf{I}_{m_n})$ .

**Remark 2.** The post-screening model sizes  $|\hat{\mathbf{S}}_k|$  in Step (ii) are random, because without condition  $(A_6)$  the selected set  $\hat{\mathbf{S}}_k$  need not coincide with the true support  $\mathbf{S}$ . Consequently, the ensuing Bahadur expansions depart from the classical fixed-dimension form, as the estimator dimension is data-dependent. Moreover, Theorem 4 shows that, in large samples, the partial-LR can be represented as a quadratic form. Specifically, it is equivalent to the squared Euclidean norm of the appropriately standardized difference between the parameter estimates obtained under the null and under the alternative. This approximation is the key tool for deriving the large-sample distributions of the refitted DR-PLR, as well as for the dimension-reduced Rao score and Wald tests.

**Corollary 2.** If  $(A_1)$ ,  $(A_2)$ ,  $(A_3)$ ,  $(A_4)$ ,  $(A_5)$ ,  $(A_6)$ , and  $(B_1)$  hold, then under  $H_{a,2}^{(n)}$ , if  $\hat{\mathbf{S}}_1 = \hat{\mathbf{S}}_2 = \hat{\mathbf{S}}$ , we have

$$T_{n,2} = \frac{1}{2}\|\zeta_{n,1} + \zeta_{n,2}\|_2^2 + o_P(m_n). \quad (15)$$

By (15) and Theorem 4(iii), for each split  $k$  the vector  $\zeta_{n,k}$  serves, up to negligible remainder, as an asymptotic surrogate for the log-partial-likelihood ratio on that split. Summing these surrogates therefore provides a full-sample summary of the likelihood evidence. In conjunction with (15), this motivates defining the refitted DR-PLR statistic for the entire dataset as

$$T_{n,\mathbf{L}} = \frac{1}{2}\|\zeta_{n,1} + \zeta_{n,2}\|_2^2, \quad (16)$$

which has the same limiting law as  $T_{n,2}$  (see Section 3.2).

By Theorem 4(iii) we have, for each split  $k$ ,

$$T_{n,\mathbf{L}}^{(k)} = \|\zeta_{n,k}\|_2^2 + o_P(m_n).$$

It follows that

$$T_{n,\mathbf{L}} = T_{n,\mathbf{L}}^* + \zeta_{n,1}^\top \zeta_{n,2},$$

where  $\zeta_{n,1}^\top \zeta_{n,2}$  serves as a correction that recoups the power loss of  $T_{n,\mathbf{L}}^*$  caused by the inflated degrees of freedom.

The choice of  $T_{n,\mathbf{L}}$  is justified on two grounds. First, combining (15) with (16) shows that  $T_{n,\mathbf{L}}$  closely approximates the statistic  $T_{n,2}$ , and in fact they have the same asymptotic distribution

(see Theorem 5). Second,  $T_{n,L}$  mitigates the spurious-selection issue by aggregating the  $\zeta_{n,k}$ 's, which encode the splitwise likelihood information, rather than relying on two independent chi-square contributions.

### 3.2 Asymptotic distributions of the refitted tests

We now characterize the null limiting law of the refitted statistics.

**Theorem 5.** Under conditions  $(A_1)$ ,  $(A_2)$ ,  $(A_3)$ ,  $(A_4)$ ,  $(A_5)$ ,  $(B_1)$ ,  $(B_2)$ , and  $(B_4)$ , suppose  $\frac{s_n+q_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), and there exist constants  $C_1, C_2 \in (0, \infty)$ , independent of  $n$ , such that  $\|\mathbf{A}\mathbf{A}^\top\|_{\text{op}} \leq C_1$  and  $\|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_{\text{op}} \leq C_2$ . Then, under  $H_{0,2}$ ,  $T \stackrel{a}{\sim} \chi_{m_n}^2$ , where  $T$  is  $T_{n,L}$ ,  $T_{n,W}$ , or  $T_{n,S}$ , which means that  $T_{n,L} \stackrel{a}{\sim} \chi_{m_n}^2$ ,  $T_{n,W} \stackrel{a}{\sim} \chi_{m_n}^2$ , and  $T_{n,S} \stackrel{a}{\sim} \chi_{m_n}^2$ .

**Remark 3.** For a test at level  $\alpha$ , reject  $H_0$  whenever  $T > \chi_{m_n, \alpha}^2$  (the upper- $\alpha$  quantile of  $\chi_{m_n}^2$ ), with  $T \in \{T_{n,L}, T_{n,W}, T_{n,S}\}$ . By Theorem 5, each refitted statistic has the same asymptotic null law as the full-sample PLR statistic  $T_{n,2}$ . A notable distinction is that the refitted procedures do *not* require condition  $(A_6)$ —a key assumption used by  $T_{n,2}$  and by certain regularization-based tests (e.g., the partial-penalized LR of Shi et al., 2019) to attain an oracle screening property. As a consequence, the refitted tests exhibit stable size and reduced tuning sensitivity across a broad range of  $\lambda_n$ .

We then establish the asymptotic alternative distribution of the refitted statistics. To give the asymptotic distributions of the refitted statistics under local alternative hypotheses, we introduce the following notational convenience. Let

$$\mathbf{I}_k(\boldsymbol{\theta}) = E \left\{ -N^{-1} \begin{pmatrix} \partial^2 Q_n(\boldsymbol{\theta}) / \partial \boldsymbol{\beta}_{\mathbf{s}^{(k)}} \partial \boldsymbol{\beta}_{\mathbf{s}^{(k)}}^\top & \partial^2 Q_n(\boldsymbol{\theta}) / \partial \boldsymbol{\beta}_{\mathbf{s}^{(k)}} \partial \boldsymbol{\gamma}^\top \\ \partial^2 Q_n(\boldsymbol{\theta}) / \partial \boldsymbol{\beta}_{\mathbf{s}^{(k)}} \partial \boldsymbol{\gamma}^\top & \partial^2 Q_n(\boldsymbol{\theta}) / \partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\top \end{pmatrix} \right\}$$

where  $k = 1, 2$ . Write the inverse matrix  $\mathbf{I}_k^{-1}(\boldsymbol{\theta})$  as  $\mathbf{I}_k^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{I}_k^{11}(\boldsymbol{\theta}) & \mathbf{I}_k^{12}(\boldsymbol{\theta}) \\ \mathbf{I}_k^{21}(\boldsymbol{\theta}) & \mathbf{I}_k^{22}(\boldsymbol{\theta}) \end{pmatrix}$  with diagonal blocks of dimension  $s_{nk} \times s_{nk}$  and  $q_n \times q_n$ , respectively. Then

$$[\mathbf{I}_k^{22}(\boldsymbol{\theta})]^{-1} = \mathbf{I}_{22}(\boldsymbol{\theta}) - \mathbf{I}_{21}(\boldsymbol{\theta}) \tilde{\mathbf{e}}^{(k)\top} [\tilde{\mathbf{e}}^{(k)} \mathbf{I}_{11}(\boldsymbol{\theta}) \tilde{\mathbf{e}}^{(k)\top}]^{-1} \tilde{\mathbf{e}}^{(k)} \mathbf{I}_{12}(\boldsymbol{\theta}).$$

Define

$$\begin{aligned} \eta_{n,L}^2 &= n \{ \boldsymbol{\Delta}_n^\top [\mathbf{A} \mathbf{I}_{r,1}^{22}(\boldsymbol{\theta}^*) \mathbf{A}^\top]^{-1} \boldsymbol{\Delta}_n + \boldsymbol{\Delta}_n^\top [\mathbf{A} \mathbf{I}_{r,2}^{22}(\boldsymbol{\theta}^*) \mathbf{A}^\top]^{-1} \boldsymbol{\Delta}_n \} / 2, \\ \eta_W^2 &= n \| [\mathbf{A} \mathbf{I}_{r,1}^{22}(\boldsymbol{\theta}^*) \mathbf{A}^\top]^{-1/2} \boldsymbol{\Delta}_n + [\mathbf{A} \mathbf{I}_{r,2}^{22}(\boldsymbol{\theta}^*) \mathbf{A}^\top]^{-1/2} \boldsymbol{\Delta}_n \|^2 / 4. \end{aligned}$$

**Theorem 6.** (*Asymptotic alternative distribution*) Suppose the conditions in Theorem 5 hold,  $\frac{s_n+q_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), and there exist constants  $C_1, C_2 \in (0, \infty)$ , independent of  $n$ , such that  $\|\mathbf{A}\mathbf{A}^\top\|_{\text{op}} \leq C_1$  and  $\|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_{\text{op}} \leq C_2$ . If  $\|\Delta_n\|_2 = O(\sqrt{m_n/n})$ , then under the alternative  $H_{a,2}^{(n)}$ ,  $T \stackrel{a}{\sim} \chi_{m_n}^2(\eta_{n,\mathbf{L}}^2)$ , where  $T$  is  $T_{n,\mathbf{L}}$ ,  $T_{n,\mathbf{W}}$ , or  $T_{n,\mathbf{S}}$ , and  $\eta_{n,\mathbf{L}}^2 = (4\phi)^{-1}n \left\| \sum_{k=1}^2 \{\mathbf{A}\mathbf{I}_k^{22}(\theta^*)\mathbf{A}^\top\}^{-1/2} \Delta_n \right\|_2^2$ .

**Remark 4.** Theorem 6 implies that the refitted procedures are consistent even without invoking condition (A<sub>6</sub>). In the boundary case  $\Delta_n = 0$ , the local alternative collapses to  $H_0$ , so the asymptotic laws under the null and under the alternative are the same. Consequently, the refitted tests attain the nominal level asymptotically and thus exhibit correct size under  $H_0$ .

**Remark 5.** Theorem 6 delivers the asymptotic local-alternative laws for the refitted procedures without invoking condition (A<sub>6</sub>). This permits a principled assessment of power across the wide range of  $\lambda_n$  values typically used in practice. If (A<sub>6</sub>) does hold, then  $P(\hat{\mathbf{S}}_k = \mathbf{S}) \rightarrow 1$  and consequently  $\eta_{n,\mathbf{L}}^2 = \eta_{n,o}^2$ . In particular,

$$T_{n,\mathbf{L}} \stackrel{a}{\sim} \chi_{m_n}^2(\eta_{n,o}^2), \quad T_{n,\mathbf{W}} \stackrel{a}{\sim} \chi_{m_n}^2(\eta_{n,o}^2), \quad T_{n,\mathbf{S}} \stackrel{a}{\sim} \chi_{m_n}^2(\eta_{n,o}^2),$$

so the refitted tests are, asymptotically, as powerful as the oracle DR-PLR benchmark.

Finally, in settings with strong covariate correlations, the refitted procedures retain appreciable power, often rivaling the oracle PLR benchmark.

**Theorem 7.** (*Nonignorable power*) Under conditions (A<sub>1</sub>)–(A<sub>5</sub>), (B<sub>1</sub>), and (B<sub>3</sub>), we have  $\eta_n^2 \leq \eta_{n,\mathbf{L}}^2$  and  $K(\rho^*)\eta_{n,o}^2 \leq \eta_{n,\mathbf{L}}^2 \leq \eta_{n,o}^2$ , with  $K(\rho^*) = \frac{1}{1+4c+4c^2}$ , where  $c = \frac{\rho^*}{\sqrt{1-\rho^*}} \sqrt{\frac{1+\rho^*}{1-\rho^*}}$  and  $\rho^* \in [0, 1)$ . In addition, if condition (A<sub>6</sub>) holds, then  $\eta_{n,\mathbf{L}}^2 = \eta_{n,o}^2$ .

**Remark 6.** One readily checks that  $K(\rho^*) \in [0, 1]$  and decreases strictly in  $\rho^*$ . Table T1 reports how  $K(\rho^*)$  evolves as  $\rho^*$  varies. This implies that, in adverse high-correlation regimes, the power of the refitted procedures is governed by  $\rho^*$ : their power is maximized at  $\rho^* = 0$  and decreases as  $\rho^*$  grows. A lower limit for  $\eta_{n,\mathbf{L}}^2$  further guarantees a nontrivial power floor for the refitted tests relative to the oracle PLR benchmark even under the most correlated scenarios. Moreover, if condition (A<sub>6</sub>) holds, the refitted tests match the oracle power regardless of the value of  $\rho^*$ .

Table T1: Change in  $K(\rho^*)$  over  $\rho^*$

| $\rho^*$    | 0.0   | 0.1   | 0.2   | 0.3   | 0.4   | 0.5   | 0.6   | 0.7   | 0.8   | 0.9   |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $K(\rho^*)$ | 1.000 | 0.658 | 0.417 | 0.256 | 0.151 | 0.084 | 0.043 | 0.020 | 0.007 | 0.002 |

## 4 Simulations

For each replication, we drew  $n = 400$  independent feature vectors  $\mathbf{W} = (w_1, \dots, w_p)^\top$  from  $\mathcal{N}_p(\mathbf{0}, \Sigma)$  with  $\Sigma_{ij} = \rho^{|i-j|}$ , where  $\rho = \frac{1}{4}$ . The whole experiment was repeated 1000 times.

The censoring time  $C$  is exponentially distributed with mean  $U * \exp\{\mathbf{W}^\top \boldsymbol{\theta}\}$ , where  $U \sim \text{Unif}(1, 3)$  was re-sampled at the beginning of each run. This design, adapted from Fan and Li (2002), yields roughly 30% censored observations in every simulated data set. We used adaptive LASSO and SCAD penalties, and the associated regularization parameters  $\lambda_n$  were selected using the Schwarz information criterion (SIC) following Shi et al. (2019).

We compared DR-PLR against a quasi-likelihood ratio benchmark computed under oracle knowledge. The oracle test assumes knowledge of the zero positions in the parameter vector  $\boldsymbol{\theta}$ , providing a benchmark for optimal performance. We investigated two dimensional settings: low-dimensional with  $r_n = p_n + q_n = 200$  and high-dimensional with  $r_n = p_n + q_n = 2000$ . We also examined two levels for the dimensionality of the parameters of interest  $\boldsymbol{\gamma}$ :  $q_n = 1$  and  $q_n = \lceil 0.5n^{1/3} \rceil = 4$ . In this paper, our discussion of the DR-PLR test mainly centers on two types of hypotheses: global-null test:

$$H_{0,1} : \boldsymbol{\gamma} = \mathbf{r} \quad \text{vs.} \quad H_{a,1} : \boldsymbol{\gamma} \neq \mathbf{r}$$

and general linear hypothesis test:

$$H_{0,2} : \mathbf{A}\boldsymbol{\gamma} - \mathbf{A}\mathbf{r} = \mathbf{0} \quad \text{vs.} \quad H_{a,2} : \mathbf{A}\boldsymbol{\gamma} - \mathbf{A}\mathbf{r} \neq \mathbf{0}$$

For each illustrative case, we estimate the null rejection rate and power for each test by running 1000 simulations. We set  $n=400$  in each simulation. Since we use a data-adaptive  $L_1$  regularizer, we must compute the associated weights  $\tilde{\boldsymbol{\theta}}$ . We obtain  $\tilde{\boldsymbol{\theta}}$  by fitting the Cox partial likelihood with an additional  $L_2$  (ridge) penalty, following the idea of Fan & Li (2002). The ridge term guarantees that every coordinate of  $\tilde{\boldsymbol{\theta}}$  is non-zero, so all adaptive weights  $\frac{1}{\tilde{\boldsymbol{\theta}}}$  are finite. For the DR-PLR test, we choose  $\lambda_n$  by minimizing SIC; unless noted otherwise, we denote the SIC-minimizing choice by  $\hat{\lambda}_{\text{opt}}$ . For the oracle quasi-likelihood ratio test, no  $\lambda_n$  selection is needed since the support of the nonzero parameters is assumed known.

We analyze two hypothesis-testing problems: (i)  $H_{0,1} : \boldsymbol{\gamma} = \mathbf{0}$  vs.  $H_{a,1} : \boldsymbol{\gamma} = 8\xi n^{-1/2}$ , where  $\boldsymbol{\gamma} = \theta_{197}$ ; and (ii)  $H_{0,2} : \mathbf{A}\boldsymbol{\gamma} = \mathbf{0}$  vs.  $H_{a,2} : \mathbf{A}\boldsymbol{\gamma} = 8\xi n^{-1/2}$ , where  $\boldsymbol{\gamma} = (\theta_{197}, \theta_{198}, \theta_{199}, \theta_{200})^\top$

and  $\mathbf{A} = (1, 1, 1, 1)^\top$ . We consider testing  $H_{0,j}$  versus a sequence of contiguous alternatives  $H_{a,j}$ , with  $\xi \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ , where  $j = 1, 2$ . When  $\xi = 0$ , the null and alternative coincide, so the null rejection rate should be approximately the significance level  $\alpha$ . As the departure parameter  $\xi$  increases, the power should increase accordingly.

In each replication, we randomly split the sample into two equal halves  $\mathcal{D}_1$  and  $\mathcal{D}_2$ . Using  $\mathcal{D}_1$ , we fit a penalized Cox model (adaptive LASSO or SCAD) with  $\lambda_n$  chosen by BIC to obtain a working active set for the nuisance part. On  $\mathcal{D}_2$ , we refit the unrestricted and restricted Cox models (the latter imposes the linear constraint through reparameterization) and compute the DR-PLR “information vector”  $\boldsymbol{\zeta}_{n,1}$ . Then, we swap the roles of the halves to obtain  $\boldsymbol{\zeta}_{n,2}$  analogously. Finally, we combine the two pieces of information to form the refitted DR-PLR statistic  $T_{n,L} = \frac{1}{2} \|\boldsymbol{\zeta}_{n,1} + \boldsymbol{\zeta}_{n,2}\|_2^2$ , which, under  $H_0$ , is asymptotically  $\chi_m^2$ , where  $\text{rank}(A) = m$ . The p-value is obtained from this reference distribution.

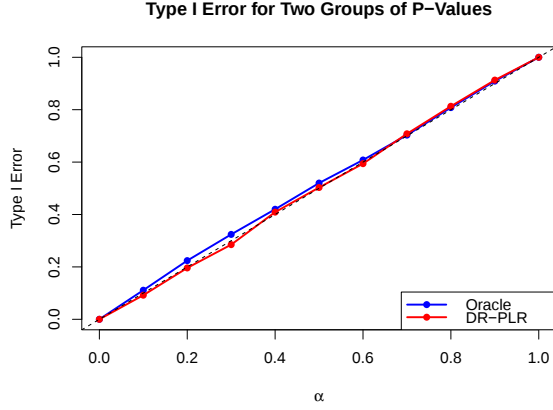
To calculate the type I error rate, we conducted 1000 simulations under  $H_0$  at each significance level  $\alpha = 0.1, 0.2, \dots, 0.9$ , and recorded the rejection count of  $H_0$ . We then treated the resulting rejection proportion as the empirical type I error probability for  $\alpha$ . Figures F1 and F2 report the empirical size under  $H_{0,1}$  and  $H_{0,2}$  against the significance level  $\alpha$  for all tests with  $\lambda_n = \hat{\lambda}_{\text{opt}}$ . Taken together, it seems that, regardless of the dimension configuration and whether the adaptive LASSO or SCAD penalty is employed, the DR-PLR test maintains the significance level, and its performance is comparable to that of the oracle test.

Table T2 reports the rejection probabilities under  $H_{a,2}$  at the nominal level  $\alpha = 0.05$  for three different levels  $\lambda_n$ :  $0.75\hat{\lambda}_{\text{opt}}$ ,  $\hat{\lambda}_{\text{opt}}$ , and  $1.25\hat{\lambda}_{\text{opt}}$ . These three levels correspond to under-shrinkage, right shrinkage, and over-shrinkage, respectively. From Table T2, we conclude that:

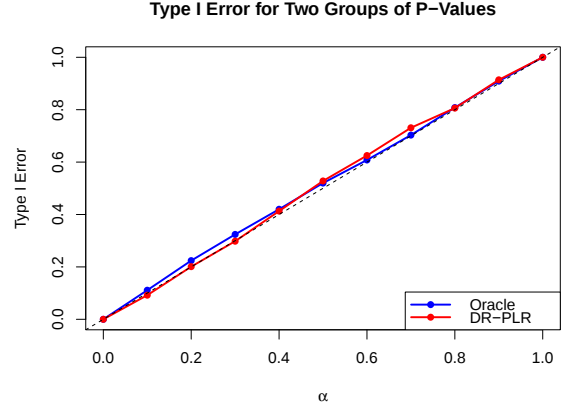
- (i) Across all dimensions ( $r_n = 200$  and  $r_n = 2000$ ) and all three tuning levels ( $0.75\hat{\lambda}_{\text{opt}}$ ,  $\hat{\lambda}_{\text{opt}}$ , and  $1.25\hat{\lambda}_{\text{opt}}$ ), the empirical power of the DR-PLR test is virtually indistinguishable from that of the oracle PLR benchmark.
- (ii) At  $\hat{\lambda}_{\text{opt}}$  or  $1.25\hat{\lambda}_{\text{opt}}$ , DR-PLR matches the oracle while maintaining the nominal 5% level, making this range the safest choice when size control is paramount.
- (iii) The  $0.75\hat{\lambda}_{\text{opt}}$  level provides the highest power—especially with SCAD—but previous size tables show a modest inflation of the null rejection rate under this tuning.

#### 4.1 Real Data Example

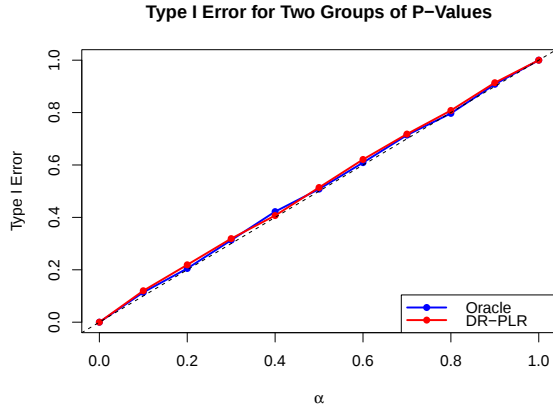
For empirical analysis, we examine a gene expression dataset consisting of 191 biopsy specimens from untreated follicular lymphoma patients—a subtype of non-Hodgkin’s lymphoma. The



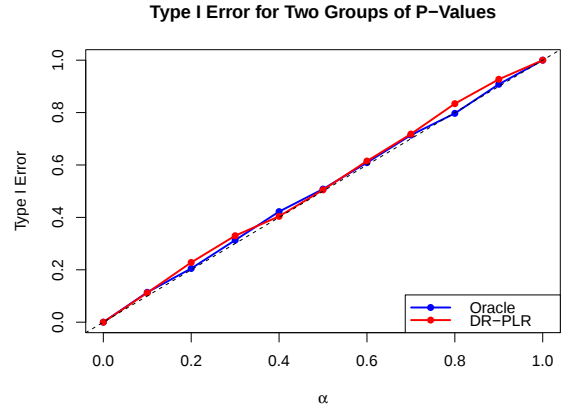
(a)  $n = 200$ , adaptive LASSO ( $q_n = 1$ )



(b)  $n = 200$ , SCAD ( $q_n = 1$ )



(c)  $n = 2000$ , adaptive LASSO ( $q_n = 1$ )

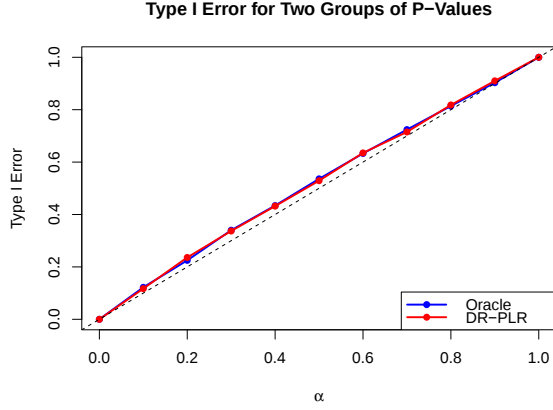


(d)  $n = 2000$ , SCAD ( $q_n = 1$ )

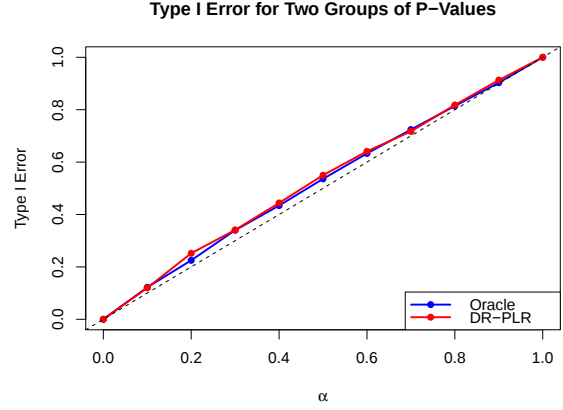
Figure F1: Plot of the null rejection rate versus  $\alpha$  under  $H_{0,1}$ . The blue curve depicts DR-PLR's null rejection rate; the red curve gives the oracle QLR counterpart.

specimens were collected at seven medical centers between 1974 and 2001. The median age of the cohort at the time of diagnosis was 51 years, and the median time was 6.6 years. Total RNA extracted from fresh-frozen tumors was profiled on Affymetrix Human Genome U133A and U133B microarrays.

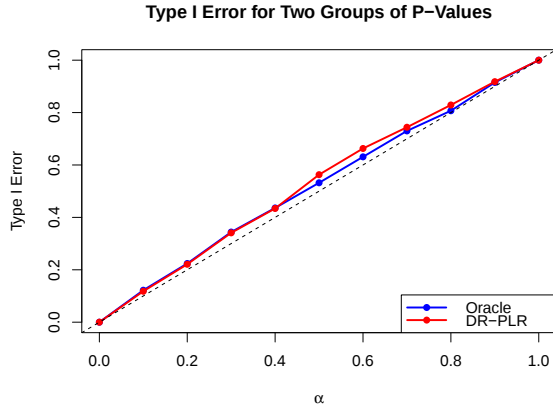
The initial expression matrix comprised 44 187 probe sets, of which 40 330 mapped to annotated genes. The signal intensities were  $\log_2$ -transformed before visualization and all subsequent analyses. For genes represented by two to seven probe sets, the median intensity was taken to obtain a single expression measure, resulting in 15 614 unique genes. The effects of these genes were treated as nuisance parameters, while five clinical characteristics—extra-nodal site, age,



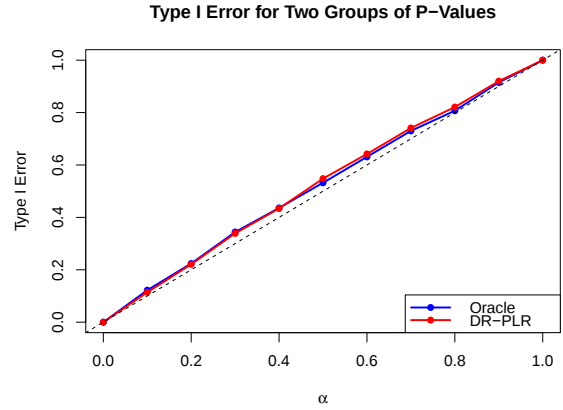
(a)  $n = 200$ , adaptive LASSO ( $q_n = 4$ )



(b)  $n = 200$ , SCAD ( $q_n = 4$ )



(c)  $n = 2000$ , adaptive LASSO ( $q_n = 4$ )



(d)  $n = 2000$ , SCAD ( $q_n = 4$ )

Figure F2: Plot of the null rejection rate versus  $\alpha$  under  $H_{0,2}$ . The blue curve depicts DR-PLR's null rejection rate; the red curve gives the oracle QLR counterpart.

LDH level, performance status (Pstat), and stage—were designated as parameters of interest. Because some samples contained missing values, we ultimately retained only 156 complete cases: 78 in the training set and 78 in the test set. On the training set, we fit a penalized Cox model (adaptive LASSO or SCAD) with  $\lambda_n$  chosen by BIC to obtain a working active set for the gene nuisance block. On the test set, we fit two Cox models in turn: an unrestricted model that includes the clinical covariates, and a restricted model that imposes the null-hypothesis constraints on the clinical covariates. Using the estimates and information matrices from the two fits, we construct the first “information vector”  $\zeta_{n,1}$ . We then swap the roles—screen on the test set and refit on the training set—to obtain  $\zeta_{n,2}$ . Finally, we combine the information

to form the statistic  $T_{n,L} = \frac{1}{2} \|\zeta_{n,1} + \zeta_{n,2}\|_2^2$ , from which the p-value is computed. Using the joint constraint on the clinical block  $A = I_5$  ( $df = 5$ ), the refitted DR-PLR test strongly rejects the global null ( $p < 0.001$ ), indicating an overall clinical signal after accounting for the screened gene set. For marginal ( $df = 1$ ) tests with  $A = e_j^\top$  in the order (Extra-nodal site, Age, LDH, Pstat, Stage), the p-values were:  $p_{\text{Extra}} = 0.149$ ,  $p_{\text{Age}} = 0.001$ ,  $p_{\text{LDH}} = 0.006$ ,  $p_{\text{Pstat}} = 0.011$ ,  $p_{\text{Stage}} = 0.093$ . Thus, Age, LDH, and Pstat remain significant after adjustment; Stage is marginal (not significant at 5% but close at 10%); and Extra-nodal site shows no independent association. These findings align with the joint rejection and underscore that the clinical block contributes to risk, with Age and LDH carrying the strongest adjusted effects.

Table T3 displays the results of the BIC method with adaptive LASSO and SCAD penalties. Both penalties yield sign-consistent coefficients for the genes jointly selected, but the resulting sparsity differs: adaptive LASSO retains 20 genes plus the five clinical covariates, while SCAD selects 13 genes plus the same covariates, producing a more compact model. Across methods, Age, LDH, performance status (Pstat), and Stage are positively associated with risk and statistically significant, while Extra-Nodal Site is included but not significant. Among the genes jointly selected by the two methods (e.g., KRT4, LOC339448, PCDH11X, SNN, SOX15, YLPM1, C19orf10, C5orf19, FLJ10815), the coefficient signs are concordant (i.e., the effect directions agree), indicating a robust determination of effect direction. Signals such as YLPM1 and LOC339448 show robust negative associations under both methods, whereas KRT4 exhibits a positive association; several additional genes are method-specific (e.g., GMPPB appears under adaptive LASSO but not SCAD), indicating some sensitivity to the choice of penalty. Overall, the two procedures agree on the dominant clinical predictors and on effect directions for many shared genes, with SCAD offering a sparser signature and adaptive LASSO capturing a broader set of candidates; method-specific findings warrant validation on independent data.

Table T2: Power (%) for Cox model under  $H_{a,2}$  at  $\alpha = 0.05$ .

| Method                 | $\lambda_n/\hat{\lambda}_{\text{opt}}$ | Test                | $r_n$ | $\xi = 0$ | 0.2  | 0.4  | 0.6  | 0.8  | 1.0  |
|------------------------|----------------------------------------|---------------------|-------|-----------|------|------|------|------|------|
| <b>DR-PLR</b>          | N/A                                    | $T^O$               | 200   | 6.7       | 11.2 | 30.6 | 56.6 | 81.5 | 94.2 |
|                        | N/A                                    | $T^O$               | 2000  | 6         | 10.6 | 31   | 58.4 | 81.2 | 93.3 |
|                        | 0.75                                   | with adaptive lasso | 200   | 7.6       | 12.1 | 32.4 | 57   | 80.3 | 92.5 |
|                        |                                        |                     | 2000  | 8         | 13.2 | 33   | 57   | 79.1 | 91.9 |
|                        |                                        | with SCAD           | 200   | 8.7       | 15.9 | 36.3 | 60.2 | 81.8 | 92.6 |
|                        |                                        |                     | 2000  | 12.9      | 20.4 | 39.9 | 64.0 | 81.9 | 93.3 |
|                        | 1.00                                   | with adaptive lasso | 200   | 6.3       | 11.8 | 29.9 | 56.1 | 80   | 93.1 |
|                        |                                        |                     | 2000  | 5.9       | 12.9 | 29.7 | 56.2 | 79.4 | 92   |
|                        |                                        | with SCAD           | 200   | 6.4       | 11.6 | 30.1 | 56.3 | 80.6 | 93.2 |
|                        |                                        |                     | 2000  | 6         | 12.3 | 29.8 | 56.7 | 79.2 | 92.2 |
|                        | 1.25                                   | with adaptive lasso | 200   | 6.3       | 11.8 | 29.8 | 56   | 80   | 93.1 |
|                        |                                        |                     | 2000  | 5.9       | 12.9 | 29.6 | 56.1 | 79.3 | 92   |
|                        |                                        | with SCAD           | 200   | 6.6       | 11.6 | 30.2 | 56.2 | 80.6 | 93.2 |
|                        |                                        |                     | 2000  | 6         | 12.2 | 29.8 | 56.7 | 79.1 | 92.2 |
| <b>Refitted DR-PLR</b> | N/A                                    | $T^O$               | 200   | 6.7       | 11.2 | 30.6 | 56.6 | 81.5 | 94.2 |
|                        | N/A                                    | $T^O$               | 2000  | 6         | 10.6 | 31   | 58.4 | 81.2 | 93.3 |
|                        | 0.75                                   | with adaptive lasso | 200   | 6.6       | 11.1 | 30.5 | 53.7 | 79   | 91.6 |
|                        |                                        |                     | 2000  | 6.5       | 12.6 | 29.1 | 54.1 | 76.4 | 90.9 |
|                        |                                        | with SCAD           | 200   | 6.5       | 12.2 | 30.7 | 54   | 77.9 | 91.9 |
|                        |                                        |                     | 2000  | 6.9       | 12.5 | 29.3 | 53.9 | 76.3 | 90.5 |
|                        | 1.00                                   | with adaptive lasso | 200   | 5.8       | 11.4 | 27.8 | 51.7 | 75.9 | 90.6 |
|                        |                                        |                     | 2000  | 7         | 11.9 | 25.8 | 49   | 71.7 | 86.1 |
|                        |                                        | with SCAD           | 200   | 6         | 10.3 | 27.3 | 51.7 | 75.9 | 90.6 |
|                        |                                        |                     | 2000  | 6.9       | 11.7 | 23.5 | 44.4 | 65.7 | 83.2 |
|                        | 1.25                                   | with adaptive lasso | 200   | 6         | 11.6 | 27.5 | 51.4 | 75.8 | 90.4 |
|                        |                                        |                     | 2000  | 7.4       | 11.7 | 25.1 | 48   | 70.1 | 85.3 |
|                        |                                        | with SCAD           | 200   | 6         | 10.3 | 27.2 | 49.8 | 74.4 | 89.3 |
|                        |                                        |                     | 2000  | 6.8       | 11.7 | 23.5 | 44.4 | 65.8 | 83.2 |

Table T3: Data estimation summary of the genes

| Gene annotation                                                                                                 | adaptive LASSO                 | SCAD                            |
|-----------------------------------------------------------------------------------------------------------------|--------------------------------|---------------------------------|
| AOF2(AL831896)                                                                                                  | ×                              | 0.4584(0.3976)                  |
| C19orf10(BQ073612)                                                                                              | 1.9157(0.6078) <sup>***</sup>  | 0.6480(0.5147)                  |
| C20orf67(AK056553)                                                                                              | 0.4425(0.6451)                 | ×                               |
| C5orf19(AK056193)                                                                                               | 0.7610(0.3789) <sup>**</sup>   | 0.5548(0.3690)                  |
| CAPZE(CAK126650)                                                                                                | 0.1609(0.7874)                 | ×                               |
| FLJ10815(BC063399)                                                                                              | 0.1037(0.4539)                 | 0.8649(0.4282) <sup>**</sup>    |
| FLJ20580(AK092734)                                                                                              | −0.8112(0.8082)                | ×                               |
| FLJ20859(CR612311)                                                                                              | −0.5694(0.6754)                | ×                               |
| FUT10(BCO63462)                                                                                                 | 0.4219(0.3749)                 | ×                               |
| GMPPB(CR621384)                                                                                                 | 1.8080(0.5788) <sup>***</sup>  | ×                               |
| GUKI(AK124677)                                                                                                  | −1.1087(0.6309) <sup>*</sup>   | ×                               |
| HERC3(D25215)                                                                                                   | ×                              | −0.3559(0.1475) <sup>**</sup>   |
| KRT4(AK056254)                                                                                                  | 0.9590(0.2723) <sup>***</sup>  | 0.7661(0.2314) <sup>***</sup>   |
| LOC339448(AK125092)                                                                                             | −1.9170(0.4793) <sup>***</sup> | −1.61302(0.4393) <sup>***</sup> |
| PCDH11X(AF332218)                                                                                               | 0.4826(0.2186) <sup>**</sup>   | 0.3406(0.2154) <sup>*</sup>     |
| POLR18(BX647683)                                                                                                | 1.4799(0.5375) <sup>***</sup>  | ×                               |
| RABL4(BX537634)                                                                                                 | −1.3685(0.5941) <sup>**</sup>  | ×                               |
| SNN(NM-003498)                                                                                                  | −1.1596(0.4687) <sup>**</sup>  | −0.9168(0.3629) <sup>**</sup>   |
| SOX15(AB006867)                                                                                                 | 0.0712(0.5503)                 | 0.0999(0.5040)                  |
| TFBIM(BU739337)                                                                                                 | −1.0437(0.5004) <sup>**</sup>  | ×                               |
| TM9SF1(BX161390)                                                                                                | −1.2249(0.5666) <sup>**</sup>  | ×                               |
| UNQ846(BC071780)                                                                                                | ×                              | 0.4623(0.2367) <sup>*</sup>     |
| XBP1(AK093842)                                                                                                  | ×                              | 0.3118(0.2028)                  |
| YLPM1(AK090435)                                                                                                 | −2.9386(0.7165) <sup>***</sup> | −2.3165(0.6005) <sup>***</sup>  |
| Extra Nodal Site                                                                                                | 0.0159(0.3758)                 | 0.3962(0.3649)                  |
| Age                                                                                                             | 0.7891(0.3285) <sup>**</sup>   | 1.1661(0.3246) <sup>***</sup>   |
| LDH                                                                                                             | 0.7833(0.2987) <sup>***</sup>  | 0.8772(0.2902) <sup>***</sup>   |
| Pstat                                                                                                           | 1.2520(0.3920) <sup>***</sup>  | 0.9276(0.3528) <sup>***</sup>   |
| Stage                                                                                                           | 1.2164(0.4244) <sup>***</sup>  | 1.3783(0.4225) <sup>***</sup>   |
| Notes: × = not selected. Superscripts <sup>***</sup> , <sup>**</sup> , <sup>*</sup> denote significance levels. |                                |                                 |

## Appendix I. Notation and conditions

First, we employ some notations to ease presentation. For a matrix  $\mathbf{C}$ , choose  $\mathbf{M}$  from its column labels  $\{1, \dots, p\}$ , Suppose  $|\mathbf{M}|$  is the count of nonzero components in  $\mathbf{M}$ , and let  $\mathbf{C}_{\mathbf{M}}$  be the submatrix of  $\mathbf{C}$  formed by column indexes in  $\mathbf{M}$ . Given  $\mathbf{C}$ , write  $\mathbf{C}_{\mathbf{M}_1, \mathbf{M}_2}$  for the submatrix with rows indexed by  $\mathbf{M}_1$  and columns indexed by  $\mathbf{M}_2$  ( $\mathbf{M}_1 \subseteq \{1, \dots, m\}$ ,  $\mathbf{M}_2 \subseteq \{1, \dots, p\}$ ). Let  $B \in \mathbb{R}^{s \times s}$  be symmetric and  $\lambda_{\max}(B)$  and  $\lambda_{\min}(B)$  are maximum and minimum eigenvalues. For a generic eigenvalue, we can use  $\lambda(B)$ . For any  $t \times 1$  vectors  $\mathbf{a}$  and  $\mathbf{b}$  with the  $i$ -th element being  $a_i$  and  $b_i$ , respectively, we let  $|\mathbf{a}| = (|a_1|, \dots, |a_t|)^\top$  and  $\mathbf{a} \circ \mathbf{b} = (a_1 b_1, \dots, a_t b_t)^\top$ . Further, we let  $\|\mathbf{a}\|_\infty = \max_j |\mathbf{a}_j|$  and  $\|\mathbf{a}\|_0$  is the count of nonzero components of  $\mathbf{a}$ . For any  $s \times t$  matrix  $\mathbf{F}$  with the  $(i, j)$ th element being  $F_{ij}$ , let  $\|\mathbf{F}\|_\infty$ ,  $\|\mathbf{F}\|_1$ , and  $\|\mathbf{F}\|_2$  be the  $L_\infty$ ,  $L_1$ , and  $L_2$  norms, respectively, defined as  $\|\mathbf{F}\|_\infty = \max_j \sum_{i=1}^s |F_{i,j}|$ ,  $\|\mathbf{F}\|_1 = \max_i \sum_{j=1}^t |F_{i,j}|$ , and  $\|\mathbf{F}\|_2 = \{\lambda_{\max}(\mathbf{F}^\top \mathbf{F})\}^{1/2}$ . Denote by  $s_{nk} = \|\hat{\mathbf{S}}_k\|_0$ , and  $d_n = 2^{-1} \min_{j \in \mathbf{S}} |\theta_j^*|$  which is the half minimum signal length of  $\beta_{\mathbf{S}}^*$ . Let  $\mathbf{W} = (\mathbf{w}_1 \dots, \mathbf{w}_n)^\top$  be the full-sample design matrix, and let  $\mathbf{P}_{\mathbf{W}} = \mathbf{W}(\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top$  be the projection matrix. Further, let  $\mathbf{W}_j$  be the  $j$ -th column of  $\mathbf{W}$ . Define  $\mathbf{I}_{\mathbf{M}}(\theta) = E(-n^{-1} \partial^2 Q_n(\theta) / \partial \theta_{\mathbf{M}} \partial \theta_{\mathbf{M}}^\top)$ .

Next, we introduce the following regularity conditions for our theorems.

(A<sub>1</sub>) For each  $\lambda_n > 0$ , the penalty  $\rho(t; \lambda_n)$  is concave and nondecreasing on  $[0, \infty)$  and is continuously differentiable with right derivative at the origin satisfying  $\rho'(0+; \lambda_n) > 0$ . Moreover, for any fixed  $t \geq 0$ , the map  $\lambda \mapsto \rho'(t; \lambda)$  is monotone on  $(0, \infty)$ , while  $\rho'(0+; \lambda)$  does not depend on  $\lambda$ .

(A<sub>2</sub>) Fix a compact neighborhood  $\mathcal{B}$  of  $\beta^*$  for which the statements below hold:

(i) For  $j$  can be 0, 1, and 2, take  $s^{(j)}$  defined on  $\mathcal{B} \times [0, \tau]$ —scalar for  $j = 0$ , vector for  $j = 1$ , and matrix for  $j = 2$ —so that the conditions below are satisfied  $\sup_{(\beta, t) \in \mathcal{B}_1 \times [0, \tau]} \|S_n^{(j)}(\beta, t) - s^{(j)}(\beta, t)\|_2 = o_p(1)$ .

(ii) On  $\mathcal{B} \times [0, \tau]$ ,  $s^{(0)}$  admits a positive lower bound and  $s^{(j)}$  are uniformly bounded. Moreover, the collection  $\{s^{(j)}(\cdot, t) : (j, t) \in \{0, 1, 2\} \times [0, \tau]\}$  share a common modulus of continuity at  $\beta^*$ .

(iii) Set  $\mathbf{e}(\beta, t) = \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)}$ ,  $\mathbf{v}(\beta, t) = \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} - \mathbf{e}(\beta, t)^{\otimes 2}$ , and  $\Sigma_{\beta}(t) = \int_0^t \mathbf{v}(\beta, u) s^{(0)}(\beta^*, u) d\Lambda_0(u)$ . Write  $\mathbf{v}(\beta_1, t)$  for the population analogue of  $\mathbf{V}(\beta_1, t)$ , obtained by the substitution  $S_n^{(\ell)} \mapsto s^{(\ell)}$ . Let

$$\Sigma_{\beta_1}(t) = \int_0^t \mathbf{v}(\beta_1, u) s^{(0)}(\beta_1^*, u) d\Lambda_0(u),$$

Where  $\Sigma_{\beta_1} = \Sigma_{\beta_1}(\tau)$ . Then, for all  $(\beta, t) \in \mathcal{B} \times [0, \tau]$ ,  $\partial s^{(0)}(\beta, t)/\partial \beta = s^{(1)}(\beta, t)$  and  $\partial s^{(1)}(\beta, t)/\partial \beta = s^{(2)}(\beta, t)$ . There exists  $c_0 > 0$  with  $\lambda_{\min}(\Sigma_{\beta_1^*}) \geq c_0$  for all  $n$ ; moreover,  $\Lambda_0(\tau) < \infty$ .

(iv) Suppose  $d_n = \sup_{t \in [0, \tau]} |S_n^{(0)}(\beta^*, t) - s^{(0)}(\beta, t)|$  and  $c_n = \sup_{t \in [0, \tau]} \|\mathbf{E}_n(\beta^*, t) - \mathbf{e}(\beta^*, t)\|_\infty$ .  $d_n = O_{\text{a.s.}}(1)$ , and  $c_n = O_{\text{a.s.}}(1)$ .

(A<sub>3</sub>) Let  $\varepsilon_{ij} = \int_0^\tau (X_{ij}(t) - e_j(\beta^*, t)) dM_i(t)$ , with  $e_j(\beta^*, t)$  the  $j$ th coordinate of  $\mathbf{e}(\beta^*, t)$ . We impose a Bernstein–Cramér moment growth condition: there is a positive value  $M$  for which, for all integers  $m \geq 2$  and all  $j$ ,  $\mathbb{E}|\varepsilon_{ij}|^m \leq \frac{m!}{2} M^{m-2} \sigma_j^2$ ,  $\sigma_j^2 = \text{Var}(\varepsilon_{ij}) < \infty$ .

(A<sub>4</sub>)  $r_\sigma(\Sigma_{\beta_1^*}) \leq C_1$ ,  $r_\sigma(\Sigma_{\beta_1^*}^{-1}) \leq C_2$  for some constants  $C_1, C_2 < \infty$ .

(A<sub>5</sub>)  $\exists C < \infty : \mathbb{E} \left[ \sup_{0 \leq t \leq \tau} Y(t) \|\mathbf{w}(t)\|_2^2 e^{\beta_1^{*T} \mathbf{w}(t)} \right] \leq C s$ .

(A<sub>6</sub>) Assume the tuning parameter  $\lambda_n$  satisfies that

$$n^{-1/2 + (\frac{1}{2}\alpha + \alpha_1 - 1)_+ + \alpha_2} = o(\lambda_n) \text{ and } \sqrt{s} \lambda_n \rho'(\beta_n^*; \lambda_n) \rightarrow 0,$$

where  $\alpha_1$  as in (A8) and  $\alpha_2 > 0$ .

(A<sub>7</sub>) Let  $\mathcal{N}_0 := \overline{B}_\infty(\beta_1^*, \beta_n^*) = \{\delta \in \mathbb{R}^s : \|\delta - \beta_1\|_\infty \leq \beta_n^*\}$ ,  $\kappa_0 := \sup_{\delta \in \mathcal{N}_0} \kappa(\rho, \delta)$ . Assume  $\beta_n^* = \omega(\sqrt{s} \{n^{-1/2} + \lambda_n \rho'(\beta_n^*)\})$ ,  $\lambda_n < \frac{\lambda_{\min}(\Sigma_{\beta_1^*})}{\kappa_0}$ .

(A<sub>8</sub>) Let  $\alpha$  be a positive value and  $0 < C < \infty$ ,

$$\sup_{0 \leq t \leq \tau} \sup_{\mathbf{v}_1 \in \mathcal{B}(\beta_1^*, \beta_n^*)} \|\tilde{\mathbf{V}}(t, \mathbf{v})\|_{2, \infty} = \min \left( C \frac{\rho'(0+)}{\rho'(\beta_n^*)}, O_p(n^{\alpha_1}) \right)$$

where  $\mathcal{B}(\beta_1^*, \beta_n^*)$  denotes the radius- $\beta_n^*$  ball in  $\mathbb{R}^s$  centered at  $\beta_1^*$ , for  $\mathbf{v} = (\mathbf{v}_1^T, \mathbf{0}^T)^T$ ,

$$\tilde{\mathbf{V}}(t, \mathbf{v}) = \frac{S_n^{(0)}(\mathbf{v}, t) S_{n21}^{(2)}(\mathbf{v}, t) - S_{n2}^{(1)}(\mathbf{v}, t) (S_{n1}^{(1)}(\mathbf{v}, t))^T}{\{S_n^{(0)}(\mathbf{v}, t)\}^2} \in \mathbb{R}^{(p-s) \times s},$$

and  $\|\tilde{\mathbf{V}}(t, \mathbf{v})\|_{2, \infty} = \max_{\|\mathbf{x}\|_2=1} \|\tilde{\mathbf{V}}(t, \mathbf{v}) \mathbf{x}\|_\infty$ .

## Appendix II. Proofs of theorems

For clarity of exposition, we first fix notation and collect a few auxiliary lemmas, and then proof these theorems.

Suppose  $\mathcal{M}_* = \{i : \beta_i^* \neq 0\}$  and  $\mathcal{M}_*^c = \{i : \beta_i^* = 0\}$ . Then we can get

$$\mathbf{U}(\beta_1, t) = \sum_{i=1}^n \int_0^t \{\mathbf{w}_i(u) - \mathbf{E}^{(1)}(\beta_1, u)\} dN_i(u),$$

with  $\mathbf{E}^{(1)}(\beta_1, u) = \mathbf{E}^{(1)}((\beta_1, \mathbf{0}), \mathbf{u})$ . Then  $\mathbf{U}(\beta_1, t) = \sum_{i=1}^n \int_0^t \{\mathbf{w}_i(u) - \mathbf{E}_n^{(1)}(\beta_1, u)\} dM_i(u)$ , since  $\sum_{i=1}^n \int_0^t \{\mathbf{w}_i(u) - \mathbf{E}_n^{(1)}(\beta_1, u)\} d\Lambda_i(u) = 0$ . Use  $\mathbf{U}(\beta_1, \tau)$  as  $\mathbf{U}(\beta_1)$ . Let  $\partial\mathbf{U}(\beta_1)$  be the derivative of  $\mathbf{U}(\beta_1)$  over  $\beta_1$ . Then

$$\begin{aligned} -N^{-1}\partial\mathbf{U}(\beta_1) &= N^{-1} \sum_{i=1}^n \int_0^\tau \mathbf{V}(\beta_1, t) dN_i(t) \\ &= \int_0^\tau \mathbf{V}(\beta_1, t) S_n^{(0)}(\beta_1^*, t) \lambda_0(t) dt + N^{-1} \int_0^\tau \mathbf{V}(\beta_1, t) d\bar{M}(t) \\ &\equiv \mathcal{I}_{\beta_1} + \mathcal{W}_{\beta_1}, \end{aligned} \tag{A.1}$$

where and thereafter  $S_n^{(\ell)}((\beta_1, \mathbf{0}), t) = S_n^{(\ell)}(\beta_1, t)$  for  $\ell = 0, 1, 2$ , and  $\mathbf{V}(\beta_1, t) = \mathbf{V}((\beta_1, \mathbf{0}), t)$ .

PROOF OF THEOREM 2.1. Let  $\hat{\beta} \in \mathbb{R}^p$  be any local maximum point of (3). Then its KKT condition reads

$$\sum_{j=1}^n \int_0^\tau \{\mathbf{X}_j(t) - \mathbf{E}_n(\hat{\beta}, t)\} dN_j(t) - n\lambda_n \mathbf{u} = \mathbf{0}, \quad \mathbf{u} \in \partial P(\hat{\beta}),$$

with  $P(\beta) = \sum_{i=1}^p \rho(|\beta_i|)$  and  $\partial P(\hat{\beta})$  is the subgradient of  $P$  at  $\hat{\beta}$ . Coordinate-wise,  $u_j = \rho'(|\hat{\beta}_j|) \text{sgn}(\hat{\beta}_j)$  ( $j \in \widehat{\mathcal{M}}^*$ ), and  $|u_i| \leq \rho'(0+)$  ( $i \in \widehat{\mathcal{M}}^c$ ), with  $\widehat{\mathcal{M}}_- = \{s : \hat{\beta}_s \neq 0\}$ .  $\square$

PROOF OF THEOREM 4.1. Let  $\hat{\beta}_1^{\mathbf{o}}$  is locally optimal for the problem of dimension  $s$ . It is enough to note that

$$\min_{\beta_1 \in \Omega_s} \lambda_{\min} \left\{ \int_0^\tau \mathbf{V}(\beta_1, t) d\bar{N}(t) \right\} > n\lambda_n \kappa(\rho, \beta_1)$$

guarantees the concavity of the penalized objective  $\mathcal{C}(\beta_1, \tau)$ . Under these conditions, the criterion admits a single global maximizer, namely  $\hat{\beta}_1^{\mathbf{o}}$   $\square$

PROOF OF THEOREM 4.6. Since  $\hat{\beta}_2 = 0$  and  $\hat{\beta}$  be a maximum point of

$$\mathcal{C}(\beta, \tau) \equiv \sum_{j=1}^n \int_0^\tau \{\beta^T \mathbf{w}_j(t) - \log(S_n^{(0)}(\beta, t))\} dN_j(t) - n \sum_{s=1}^p p_{\lambda_n}(|\beta_s|), \tag{A.2}$$

$\hat{\beta}_1$  is such that

$$\mathbf{U}(\hat{\beta}_1) - n\lambda_n \rho'(|\hat{\beta}_1|) \circ \text{sgn}(\hat{\beta}_1) = 0, \tag{A.3}$$

for brevity, write  $\mathbf{U}(\beta_1) = \mathbf{U}(\beta_1, \tau)$ . By Taylor expansion, we can get

$$\mathbf{U}(\hat{\beta}_1) = \mathbf{U}(\beta_1^*) + \partial\mathbf{U}(\beta_1^*)(\hat{\beta}_1 - \beta_1^*) + \mathbf{r}_n(\tilde{\beta}_1), \tag{A.4}$$

with  $\tilde{\beta}_1 \in [\beta_1^*, \hat{\beta}_1]$ , and  $r_n(\beta_1)$  equals

$$\frac{1}{2} \sum_{j,k} (\beta_{1j} - \beta_{1j}^*)(\beta_{1k} - \beta_{1k}^*) \frac{\partial^2 \mathbf{U}(\beta_1)}{\partial \beta_{1j} \partial \beta_{1k}}.$$

Combine (A.3) and (A.4), we can get

$$\begin{aligned} \sqrt{n}(\hat{\beta}_1 - \beta_1^*) &= (-N^{-1} \partial \mathbf{U}(\beta_1^*))^{-1} n^{-1/2} \mathbf{U}(\beta_1^*) \\ &+ (-N^{-1} \partial \mathbf{U}(\beta_1^*))^{-1} n^{-1/2} \mathbf{r}_n(\tilde{\beta}_1) \\ &- (-N^{-1} \partial \mathbf{U}(\beta_1^*))^{-1} n^{1/2} \lambda_n \rho'(|\hat{\beta}_1|) \circ \text{sgn}(\hat{\beta}_1). \end{aligned} \quad (\text{A.5})$$

It can be shown that  $\|N^{-1} \frac{\partial^2 \mathbf{U}(\beta_1)}{\partial \beta_{1j} \partial \beta_{1k}}\|_2 = O_p(\sqrt{s})$  for  $\beta_1 \in \mathcal{B}(\beta_1^*, \beta_n^*)$ , hence

$$\|N^{-1} \mathbf{r}_n(\tilde{\beta}_1)\|_2 = O_p(\sqrt{s} \|\hat{\beta}_1 - \beta_1^*\|_2^2).$$

Using (15),  $\|\tilde{\beta}_1 - \beta_1^*\|_2 \leq \|\hat{\beta}_1 - \beta_1^*\|_2 = O_p(\sqrt{s/n} + \sqrt{s} \lambda_n \rho'(\beta_n^*))$ , thus

$$\|N^{-1} \mathbf{r}_n(\tilde{\beta}_1)\|_2 = O_p\{\sqrt{s}(\sqrt{s/n} + \sqrt{s} \lambda_n \rho'(\beta_n^*))^2\}.$$

We note that  $-N^{-1} \partial \mathbf{U}(\beta_1^*) = \mathcal{I}_{\beta_1^*} + \mathcal{W}_{\beta_1^*}$ . Combining Lemmas 4.1 and 3 yields that for every  $\mathbf{b}_n \in \mathbb{R}^s$  with  $\|\mathbf{b}_n\|_2 = 1$ , if  $\frac{s}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ),

$$\begin{aligned} &|\mathbf{b}_n^T \Sigma_{\beta_1^*}^{1/2} (-N^{-1} \partial \mathbf{U}(\beta_1^*))^{-1} n^{-1/2} r_n(\tilde{\beta}_1)| \\ &\leq n^{1/2} \|\Sigma_{\beta_1^*}^{1/2}\|_2 \cdot \| -N^{-1} \partial \mathbf{U}(\beta_1^*) \|_2 O_p(s\sqrt{s/n} + s\sqrt{s} \lambda_n^2 \rho'(\beta_n^*)^2) \\ &= n^{1/2} O_p(1 + s/\sqrt{n}) O_p(s\sqrt{s/n} + s\sqrt{s} \lambda_n^2 \rho'(\beta_n^*)^2) = o_p(1) \end{aligned}$$

Using the conditions stated in Theorem 6.4 we can get

$$\begin{aligned} &|\mathbf{b}_n^T \Sigma_{\beta_1^*}^{1/2} (-N^{-1} \partial \mathbf{U}(\beta_1^*))^{-1} \sqrt{n} \lambda_n \rho'(|\hat{\beta}_1|) \circ \text{sgn}(\hat{\beta}_1)| \\ &\leq \|\Sigma_{\beta_1^*}^{1/2}\|_2 \cdot \| -N^{-1} \partial \mathbf{U}(\beta_1^*) \|_2 \sqrt{sn} \lambda_n |\rho'(\beta_n^*)| \\ &= O_p(1 + s/\sqrt{n}) O_p(\sqrt{sn} \lambda_n |\rho'(\beta_n^*)|) = o_p(1), \end{aligned}$$

it demonstrates that the latter two terms in the right-hand side of (A.5), when pre-multiplied by  $\mathbf{b}_n^T \Sigma_{\beta_1^*}^{0.5}$ , are  $o_p(1)$ . Combining this and (A.5), we can get

$$\begin{aligned} \sqrt{n} \mathbf{b}_n^T \Sigma_{\beta_1^*}^{0.5} (\hat{\beta}_1 - \beta_1^*) &= \mathbf{b}_n^T \Sigma_{\beta_1^*}^{0.5} (-N^{-1} \partial \mathbf{U}(\beta_1^*))^{-1} n^{-0.5} \mathbf{U}(\beta_1^*) + o_p(1) \\ &\equiv \phi_n + o_p(1). \end{aligned}$$

By using Lemma 4.2, we can get

$$\sqrt{n} \mathbf{b}_n^T \Sigma_{\beta_1^*}^{0.5} (\hat{\beta}_1 - \beta_1^*) = \phi_{n1} + o_p(1).$$

We only need to show  $\phi_{n1}$  is asymptotically Gaussian; consequently,  $\hat{\beta}_1$  is asymptotically normal by Slutsky's theorem. For this purpose, we define  $\mathcal{I}_{\beta_1^*}(t) = \int_0^t \mathbf{V}(\beta_1^*, u) S_n^{(0)}(\beta_1^*, u) \lambda_0(u) du$  and

$$\phi_{n1}(t) = \mathbf{b}_n^T \Sigma_{\beta_1^*}^{-1/2} n^{-1/2} \mathbf{U}(\beta_1^*, t).$$

Let  $\mathcal{I}_{\beta_1^*}(\tau) = \mathcal{I}_{\beta_1^*}$ ,  $\phi_{n1}(\tau) = \phi_{n1}$ , where  $\phi_{n1}$  is a centered martingale w.r.t.  $(\mathcal{F}_t)$ , whose previsible quadratic variation is

$$\begin{aligned} \langle \phi_{n1}(t) \rangle &= \int_0^t \mathbf{b}_n^T \Sigma_{\beta_1^*}^{-1/2} N^{-1} \sum_{i=1}^n \left( \mathbf{S}_i(u) - \mathbf{E}_n^{(1)}(\beta^*, u) \right)^{\otimes 2} d\langle M_i(u) \rangle \\ &= \int_0^t \mathbf{b}_n^T \Sigma_{\beta_1^*}^{-1/2} \mathbf{V}(\beta^*, u) S_n^{(0)}(\beta^*, u) \lambda_0(u) \Sigma_{\beta_1^*}^{-1/2} \mathbf{b}_n du \\ &= \mathbf{b}_n^T \Sigma_{\beta_1^*}^{-1/2} \mathcal{I}_{\beta_1^*}(t) \Sigma_{\beta_1^*}^{-1/2} \mathbf{b}_n. \end{aligned}$$

With  $\|\mathbf{b}_n\|_2 = 1$ . According to Lemma 4.1,  $\langle \phi_{n1}(\tau) \rangle = \mathbf{b}_n^T \Sigma_{\beta_1^*}^{-1/2} \mathcal{I}_{\beta_1^*} \Sigma_{\beta_1^*}^{-1/2} \mathbf{b}_n \xrightarrow{p} 1$ . Applying the martingale CLT [?],  $\phi_{n1}$  is asymptotically  $N(0, 1)$ . □

## Appendix II. Proofs of theorems

**Proof of Theorem 1.** Our idea is to derive from the Karush-Kuhn-Tucker (KKT) equations an upper bound for  $\|\hat{\mathbf{N}}\|_0$ . Our proof consists of three steps: (i) We prove that  $\|(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)_{\mathbf{T}}\|_2 = O_P\{\sqrt{(s_n + q_n)/n}\}$  and  $\|(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)_{\mathbf{T}^c}\|_1 = O_P\{t_n/(n\lambda_n)\}$ , where  $t_n = \min\{\log r_n, s_n + q_n\}$ ; (ii) To build the inequality,  $\|\boldsymbol{\rho}(\tilde{\boldsymbol{\theta}}_{\hat{\mathbf{N}}}, \lambda_n)\|_2 \geq c_1 \sqrt{\|\hat{\mathbf{N}}\|_0}$ ; (iii) To establish the desired bound in the theorem; (iv) To show that, under condition (A<sub>6</sub>),  $P(\hat{\mathbf{S}} = \mathbf{S}) \rightarrow 1$ .

Step (i). Let  $\boldsymbol{\delta} = \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$  and  $r_n = p_n + q_n$ . By Lemma 4(ii), we have

$$\|\boldsymbol{\delta}_{\mathbf{T}}\|_2 = O_P\{\sqrt{(s_n + q_n)/n}\} \quad \text{and} \quad \|\boldsymbol{\delta}_{\mathbf{T}^c}\|_1 = O_P\{t_n/(n\lambda_n)\}, \quad (\text{A.6})$$

where  $t_n = \min\{\log r_n, s_n + q_n\}$ .

Step (ii). Since  $\tilde{\boldsymbol{\theta}}$  maximizes the objective function in (8), by the KKT conditions, we have

$$\mathbf{U}(\tilde{\boldsymbol{\theta}}) - n\lambda_n \mathbf{h} = \mathbf{0}, \quad (\text{A.7})$$

where  $h_i = 0$  for  $i \in \mathbf{D}$ ,  $h_i = \rho'_{\lambda_n}(|\tilde{\theta}_i|) \text{sgn}(\tilde{\theta}_i)$  for  $i \in \hat{\mathbf{S}}$ , and  $h_i \in [-\rho'(0+), \rho'(0+)]$  for  $i \in \hat{\mathbf{T}}^c$ . This implies that

$$\mathbf{U}_{\hat{\mathbf{S}}}(\tilde{\boldsymbol{\theta}}) - n\lambda_n \boldsymbol{\rho}(\tilde{\boldsymbol{\theta}}_{\hat{\mathbf{S}}}, \lambda_n) = \mathbf{0},$$

where  $\boldsymbol{\rho}(\tilde{\boldsymbol{\theta}}_{\hat{\mathbf{S}}}, \lambda_n) = \{\rho'_{\lambda_n}(|\tilde{\theta}_i|)\text{sgn}(\tilde{\theta}_i), i \in \hat{\mathbf{S}}\}^\top$  collects the entries of  $\mathbf{h}$  whose indices lie in  $\hat{\mathbf{S}}$ . In the same way, we can use  $\boldsymbol{\rho}(\tilde{\boldsymbol{\theta}}_{\hat{\mathbf{N}}}, \lambda_n)$  to collect the entries of  $\mathbf{h}$  whose indices lie in  $\hat{\mathbf{N}}$ . Note that  $\hat{\mathbf{N}} = \hat{\mathbf{S}} \setminus \mathbf{S}$ . Combining subadditivity of the norm with the mean-value theorem, we get

$$\begin{aligned} \lambda_n \|\boldsymbol{\rho}(\tilde{\boldsymbol{\theta}}_{\hat{\mathbf{N}}}, \lambda_n)\|_2 &= \|\mathbf{U}_{\hat{\mathbf{N}}}(\tilde{\boldsymbol{\theta}})\|_2 \\ &\leq \|\mathbf{U}_{\hat{\mathbf{N}}}(\boldsymbol{\theta}^*)\|_2 + \|\mathbf{U}_{\hat{\mathbf{N}}}(\tilde{\boldsymbol{\theta}}) - \mathbf{U}_{\hat{\mathbf{N}}}(\boldsymbol{\theta}^*)\|_2, \\ &= \|\mathbf{U}_{\hat{\mathbf{N}}}(\boldsymbol{\theta}^*)\|_2 + \|\mathbf{H}_{\hat{\mathbf{N}}, \hat{\mathbf{N}}}(\boldsymbol{\theta}_*)\boldsymbol{\delta}\|_2, \end{aligned} \quad (\text{A.8})$$

where  $\boldsymbol{\theta}_*$  lies between  $\boldsymbol{\theta}^*$  and  $\tilde{\boldsymbol{\theta}}$ . An application of Lemma 4(i) leads to

$$\|\mathbf{U}_{\hat{\mathbf{N}}}(\tilde{\boldsymbol{\theta}})\|_\infty \leq \|\mathbf{U}(\tilde{\boldsymbol{\theta}})\|_\infty = O_P(\sqrt{\log(r_n)/n}),$$

which, combined with  $\lambda_n^{-1}\sqrt{\log(r_n)/n} \rightarrow 0$ , yields that

$$P(\|\mathbf{U}_{\hat{\mathbf{N}}}(\tilde{\boldsymbol{\theta}})\|_\infty \leq c_1 \lambda_n / 2) \rightarrow 1, \quad (\text{A.9})$$

where  $c_1$  is defined in condition (A<sub>1</sub>). For any  $j \in \hat{\mathbf{N}}$ , (A.6) ensures that

$$|\tilde{\theta}_j| = |\tilde{\theta}_j - \theta_j^*| \leq \|\boldsymbol{\delta}_{\mathbf{T}^c}\|_1 = O_P\{t_n/(n\lambda_n)\},$$

which, combined with  $t_n = \min\{\log r_n, s_n + q_n\}$  and  $\lambda_n^{-1}\sqrt{\log(r_n)/n} \rightarrow 0$ , leads to

$$P(|\tilde{\theta}_j| \leq \lambda_n) \rightarrow 1, \quad j \in \hat{\mathbf{N}}.$$

By  $\inf_{t \in [0, \lambda_n]} \rho(t, \lambda_n) \geq c_1$  in condition (A<sub>1</sub>), we establish that

$$\|\boldsymbol{\rho}(\tilde{\boldsymbol{\theta}}_{\hat{\mathbf{N}}}, \lambda_n)\|_2 = \sqrt{\sum_{j \in \hat{\mathbf{N}}} [\rho'_{\lambda_n}(|\tilde{\theta}_j|)]^2}, \quad \Pr\left(\|\boldsymbol{\rho}(\tilde{\boldsymbol{\theta}}_{\hat{\mathbf{N}}}, \lambda_n)\|_2 \geq c_1 \sqrt{|\hat{\mathbf{N}}|}\right) \rightarrow 1. \quad (\text{A.10})$$

Step (iii). It follows from (A.8)-(A.10) and Hölder's inequality that, w.p. 1-o(1),

$$\begin{aligned} c_1 \lambda_n \sqrt{\|\hat{\mathbf{N}}\|_0} &\leq \|\mathbf{U}_{\hat{\mathbf{N}}}(\boldsymbol{\theta}^*)\|_2 + \|\mathbf{H}_{\hat{\mathbf{N}}, \hat{\mathbf{N}}}(\boldsymbol{\theta}_*)\boldsymbol{\delta}\|_2 \\ &\leq \sqrt{\|\hat{\mathbf{N}}\|_0} \|\mathbf{U}_{\hat{\mathbf{N}}}(\boldsymbol{\theta}^*)\|_\infty + \|\mathbf{H}_{\hat{\mathbf{N}}, \hat{\mathbf{N}}}(\boldsymbol{\theta}_*)\boldsymbol{\delta}\|_2 \\ &\leq 0.5c_1 \lambda_n \sqrt{\|\hat{\mathbf{N}}\|_0} + \sqrt{\|\mathbf{H}_{\hat{\mathbf{N}}, \hat{\mathbf{N}}}(\boldsymbol{\theta}_*)\|_2 \cdot \|\boldsymbol{\delta}^\top \mathbf{H}_{\hat{\mathbf{N}}, \hat{\mathbf{N}}}(\boldsymbol{\theta}_*)\boldsymbol{\delta}\|_2}, \end{aligned}$$

which combined with condition (A<sub>3</sub>) yields that

$$\sqrt{\|\hat{\mathbf{N}}\|_0} \leq 2(\rho_1^{1/2} c_1 \lambda_n)^{-1} \sqrt{\|\mathbf{H}_{\hat{\mathbf{N}}, \hat{\mathbf{N}}}(\boldsymbol{\theta}_*)\|_2 \cdot \{\|\boldsymbol{\delta}_{\mathbf{T}}\|_2^2 + \|\boldsymbol{\delta}_{\mathbf{T}^c}^c\|_1^2\}}. \quad (\text{A.11})$$

Applying the mean value theorem, then we can get

$$\mathbf{v}_2(\boldsymbol{\theta}_*) = \mathbf{v}_2(\boldsymbol{\theta}^*) + \text{diag}[\mathbf{v}_3(\tilde{\boldsymbol{\theta}}^*) \circ \{\mathbf{W}\tilde{\boldsymbol{\delta}}\}],$$

with  $\tilde{\boldsymbol{\theta}}^* \in [\boldsymbol{\theta}_*, \boldsymbol{\theta}^*]$  and  $\tilde{\boldsymbol{\delta}} = \tilde{\boldsymbol{\theta}}_* - \boldsymbol{\theta}^*$ . Thus, by  $\mathbf{W}\tilde{\boldsymbol{\delta}} = \mathbf{W}_{\mathbf{T}}\tilde{\boldsymbol{\delta}}_{\mathbf{T}} + \mathbf{W}_{\mathbf{T}^c}\tilde{\boldsymbol{\delta}}_{\mathbf{T}^c}$  and the definitions of  $\varphi_1(\cdot)$ ,  $\varphi_2(\cdot)$  and  $\varphi_3(\cdot)$  in condition (A<sub>4</sub>), we get

$$\begin{aligned}\|\mathbf{H}_{\hat{\mathbf{N}}, \hat{\mathbf{N}}}(\boldsymbol{\theta}_*)\|_2 &\leq \|\mathbf{H}_{\hat{\mathbf{N}}, \hat{\mathbf{N}}}(\boldsymbol{\theta}^*)\|_2 + \|N^{-1}\mathbf{W}_{\hat{\mathbf{N}}}^\top \text{diag}[\mathbf{v}_3(\tilde{\boldsymbol{\theta}}^*) \circ \{\mathbf{W}\tilde{\boldsymbol{\delta}}\}]\mathbf{W}_{\hat{\mathbf{N}}}\|_2 \\ &\leq \varphi_1(\|\hat{\mathbf{N}}\|_0) + \varphi_2(\|\hat{\mathbf{N}}\|_0)\|\tilde{\boldsymbol{\delta}}_{\mathbf{T}}\|_1 + \varphi_3(\|\hat{\mathbf{N}}\|_0)\|\tilde{\boldsymbol{\delta}}_{\mathbf{T}^c}\|_1.\end{aligned}$$

This, combined with (A.6), (A.11),  $\max\{\sqrt{(s_n + q_n)/n}, \sqrt{\log(p_n + q_n)/n}\} = o(\lambda_n)$  in Condition (A<sub>1</sub>), yields that

$$\begin{aligned}\sqrt{\|\hat{\mathbf{N}}\|_0} &\leq 2(\rho_1^{1/2}c_1)^{-1}\{\varphi_1(\|\hat{\mathbf{N}}\|_0) + \varphi_2(\|\hat{\mathbf{N}}\|_0)\|\tilde{\boldsymbol{\delta}}_{\mathbf{T}}\|_1 + \varphi_3(\|\hat{\mathbf{N}}\|_0)\|\tilde{\boldsymbol{\delta}}_{\mathbf{T}^c}\|_1\}^{1/2}\{\lambda_N^{-1}(\|\boldsymbol{\delta}_{\mathbf{T}}\|_2 + \|\boldsymbol{\delta}_{\mathbf{T}^c}\|_1)\} \\ &\leq (0.5\rho_1^{-1})^{1/2}\{\varphi_1(\|\hat{\mathbf{N}}\|_0) + \varphi_2(\|\hat{\mathbf{N}}\|_0)(s_n + q_n)\sqrt{c_n/n} + \sqrt{t_n/n}\varphi_3(\|\hat{\mathbf{N}}\|_0)\}^{1/2}\end{aligned}\quad (\text{A.12})$$

holds with probability going to one, where the last equality follows from  $\|\tilde{\boldsymbol{\delta}}_{\mathbf{T}}\|_1 \leq (s_n + q_n)\|\boldsymbol{\delta}_{\mathbf{T}}\|_2 = o_P((s_n + q_n)\sqrt{c_n/n})$  and  $\|\tilde{\boldsymbol{\delta}}_{\mathbf{T}^c}\|_1 \leq \|\boldsymbol{\delta}_{\mathbf{T}^c}\|_1 = o_P(\frac{t_n}{n\lambda_n})$  in (A.6) for  $c_n \rightarrow \infty$ . This provides an upper bound of  $\|\hat{\mathbf{N}}\|_0$ . But this upper bound is not an ideal one. Next, we derive an ideal upper bound, which independent of  $\|\hat{\mathbf{N}}\|_0$ . Define the set

$$\mathcal{M} = \{m \in \mathbb{N} \mid m \geq \rho_1^{-1}\{\varphi_1(m) + \varphi_2(m)(s_n + q_n)\sqrt{c_n/n} + \sqrt{t_n/n}\varphi_3(m)\}\},$$

and let  $m = 3\rho_2(s_n + q_n)/\rho_1$ . Using condition (A<sub>4</sub>), we get

$$\varphi_1(m) + \varphi_2(m)(s_n + q_n)\sqrt{c_n/n} + \varphi_3(m)\sqrt{t_n/n} \leq 3\rho_2(s_n + q_n). \quad (\text{A.13})$$

This implies that  $3\rho_2(s_n + q_n)/\rho_1 \in \mathcal{M}$ , i.e.  $\mathcal{M}$  is well defined and not empty. We will show that

$$\|\hat{\mathbf{N}}\|_0 \leq 0.5\rho_1^{-1} \min_{m \in \mathcal{M}} \{\varphi_1(m) + \varphi_2(m)(s_n + q_n)\sqrt{c_n/n} + \sqrt{t_n/n}\varphi_3(m)\}. \quad (\text{A.14})$$

In fact, for any  $M \in \mathcal{M}$ , if  $\|\hat{\mathbf{N}}\|_0 > M$ , then, by (A.12),

$$\begin{aligned}\|\hat{\mathbf{N}}\|_0 &\leq 0.5\rho_1^{-1}\{\varphi_1(\|\hat{\mathbf{N}}\|_0) + \varphi_2(\|\hat{\mathbf{N}}\|_0)(s_n + q_n)\sqrt{c_n/n} + \sqrt{t_n/n}\varphi_3(\|\hat{\mathbf{N}}\|_0)\} \\ &= 0.5\rho_1^{-1}\{\varphi_1(\frac{\|\hat{\mathbf{N}}\|_0}{M}M) + \varphi_2(\frac{\|\hat{\mathbf{N}}\|_0}{M}M)(s_n + q_n)\sqrt{c_n/n} + \sqrt{t_n/n}\varphi_3(\frac{\|\hat{\mathbf{N}}\|_0}{M}M)\} \\ &\leq 0.5\rho_1^{-1}\{\varphi_1(M) + \varphi_2(M)(s_n + q_n)\sqrt{c_n/n} + \sqrt{t_n/n}\varphi_3(M)\}\lceil \|\hat{\mathbf{N}}\|_0/M \rceil,\end{aligned}$$

where we can get the final step from Lemma 3 and  $\lceil a \rceil$  represents the minimum integer greater than  $a$ . Since  $\lceil \frac{\|\hat{\mathbf{N}}\|_0}{M} \rceil < 2\frac{\|\hat{\mathbf{N}}\|_0}{M}$ , we have

$$\|\hat{\mathbf{N}}\|_0 < \rho_1^{-1}[\varphi_1(M) + \varphi_2(M)(s_n + q_n)\sqrt{c_n/n} + \sqrt{t_n/n}\varphi_3(M)]\frac{\|\hat{\mathbf{N}}\|_0}{M},$$

which implies that

$$M < \rho_1^{-1} [\varphi_1(M) + \varphi_2(M)(s_n + q_n)\sqrt{c_n/n} + \sqrt{t_n/n}\varphi_3(M)].$$

This is in contradiction with  $M \in \mathcal{M}$ . Thus,  $\|\hat{\mathbf{N}}\|_0 \leq M$  holds for any  $M \in \mathcal{M}$ , which combined with functions  $\varphi_j(\cdot)$  being increasing leads to (A.14). It follows from (A.13) and (A.14) that  $\|\hat{\mathbf{N}}\|_0 \leq 1.5\rho_1^{-1}\rho_2(s_n + q_n)$ . Thus,  $P\{\|\hat{\mathbf{N}}\|_0 \leq c_0 s_n\} \rightarrow 1$ , where  $c_0 = 1.5\rho_1^{-1}\rho_2$ .

Step (iv). Under Condition (A<sub>6</sub>), we can show that  $P(\hat{\mathbf{S}} = \mathbf{S}) \rightarrow 1$ , following the argument of Theorem 2.1 in Shi et al. (2019). Then the result holds. To this end, it suffices to verify that their conditions are satisfied, that is,

- (1) *there is a local maximum point  $\tilde{\boldsymbol{\theta}}$  of  $Q_n(\boldsymbol{\theta}) - n \sum_{j=1}^{p_n} p_{\lambda_n}(\boldsymbol{\theta})$  with the constraint  $\tilde{\boldsymbol{\theta}}_{\mathbf{T}^c} = \mathbf{0}$  such that  $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = O_P(\sqrt{(s_n + q_n)/n})$ ;*
- (2)  *$\tilde{\boldsymbol{\theta}}$  is indeed a local maximizer  $Q_n(\boldsymbol{\theta}) - n \sum_{j=1}^{p_n} p_{\lambda_n}(\boldsymbol{\theta})$ .*

Similar to proof of Lemma 4, we have result (1). To prove result (2), by Theorem 1 in Fan and Lv (2011), it suffices to show that

$$\|\mathbf{U}_{\mathbf{T}^c}(\tilde{\boldsymbol{\theta}})\|_\infty < n\lambda_n\rho'(0+),$$

which can be proved by adapting the argument for Theorem 2.1 in Shi et al. (2019). □

**Proofs of Theorems 2 and 3.** For convenience, the proofs are put behind the argument for Theorem 6. □

**Proof of Corollary 1.** Using Lemma 13(i) that the result holds. In particular, if there is no nuisance block or the cross-information between  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$  vanishes, we note that the equality holds. □

**Proof of Theorem 4.**

- (i). The sparsity results are proven in Lemma 8.
- (ii). Result (a) is obtained in Lemma 8(i), and Result (b) is straightforward from Lemma 8(ii).
- (iii). Applying Lemma 11(ii),(iii) leads to

$$2\{Q_n^{(k)}(\hat{\boldsymbol{\theta}}_k) - Q_n^{(k)}(\hat{\boldsymbol{\theta}}_{0,k})\} = \|\zeta_{n,k}\|_2^2 + o_P(m_n),$$

Using Lemma 11(iii) with  $\boldsymbol{\Delta}_n = \mathbf{0}$ , we establish that

$$\zeta_{n,k} = N^{-1/2}\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + o_P(1) = \zeta_{n,k}^* + o_P(1).$$

Denoted by  $\tilde{\mathbf{A}}_o = (\mathbf{0}_{m_n \times s_n}, \mathbf{A})$  and  $\mathbf{\Omega}_o(\boldsymbol{\theta}^*) = \{\tilde{\mathbf{A}}_o \mathbf{I}_o^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_o^\top\}^{-1}$ . For convenience, by the definition of  $\mathbf{W}_T$ , let  $\mathbf{W}_T = (\mathbf{w}_{1,T}^{(k)}, \dots, \mathbf{w}_{N,T}^{(k)})^\top$ . Under condition  $(A_6)$ ,  $P(\hat{\mathbf{S}} = \mathbf{S}) \rightarrow 1$ . Then

$$\zeta_{n,k}^* = \sum_{i=1}^N N^{-1/2} \mathbf{\Omega}_o^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_o \mathbf{I}_o^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_T^{(k)}(\boldsymbol{\theta}^*) \equiv \sum_{i=1}^N \eta_{i,k}^*,$$

which be a sum of iid RVs for which  $E\eta_{i,k}^* = \mathbf{0}$  and  $E\eta_{i,k}^* (\eta_{i,k}^*)^\top = N^{-1} \phi \mathbf{I}_{m_n}$ . Applying the Cauchy-Schwarz inequality, condition  $(B_2)(i)$ , and  $\|\mathbf{\Omega}_o^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_o \mathbf{I}_o^{-1/2}(\boldsymbol{\theta}^*)\|_2^2 \leq 1$ , then we can get

$$\begin{aligned} m_n^{1/4} \sum_{i=1}^N E(\|\eta_{i,k}^*\|_2^3) &= m_n^{1/4} \sum_{i=1}^N E\|N^{-1/2} \mathbf{\Omega}_o^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_o \mathbf{I}_o^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_T^{(k)}(\boldsymbol{\theta}^*)\|_2^3 \\ &\leq m_n^{1/4} N^{-3/2} \sum_{i=1}^N E\|\mathbf{I}_o^{-1/2}(\boldsymbol{\theta}^*) \mathbf{U}_T^{(k)}(\boldsymbol{\theta}^*)\|_2^3 \rightarrow 0. \end{aligned}$$

This, combined with Theorem 1 of Bentkus (2004), yields that

$$\sup_{\boldsymbol{\psi}} \left| P\left(\sum_{i=1}^n \eta_{i,k}^* \in \boldsymbol{\psi}\right) - P(\mathcal{Z} \in \boldsymbol{\psi}) \right| \leq c_* m_n^{1/4} \sum_{i=1}^n E\|\eta_{i,k}^*\|_2^3 \rightarrow 0,$$

with  $\mathcal{Z}$  be a Gaussian random vector for which  $E(\mathcal{Z}) = \mathbf{0}$  and  $\text{Cov}(\mathcal{Z}) = \phi \mathbf{I}_{m_n}$ , and the supremum is taken for all convex subsets in  $R^{m_n}$ .  $\square$

**Proof of Corollary 2.** It follows from Lemma 11(iv) that

$$T_{n,2} = \left\| \{\tilde{\mathbf{A}}_* \mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_*^\top\}^{-1/2} \{n^{-1/2} \tilde{\mathbf{A}}_* \mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}}^{(k)}(\boldsymbol{\theta}^*) + n^{1/2} \boldsymbol{\Delta}_n\} \right\|_2^2 + o_P(m_n),$$

where  $\tilde{\mathbf{A}}_* = (\mathbf{0}_{m_n \times \|\hat{\mathbf{S}}\|_0}, \mathbf{A})$ . Under condition  $(A_6)$ , we have  $P(\hat{\mathbf{S}}_1 = \hat{\mathbf{S}}_2 = \hat{\mathbf{S}}) \rightarrow 1$ . Then  $\mathbf{U}_T(\boldsymbol{\theta}^*) = \sum_{k=1}^2 \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)$ , and

$$\{\tilde{\mathbf{A}}_* \mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_*^\top\}^{-1/2} \tilde{\mathbf{A}}_* \mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\boldsymbol{\theta}^*) = \mathbf{\Omega}_k^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*),$$

where  $\mathbf{\Omega}_k(\boldsymbol{\theta}^*) = \{\tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k^\top\}^{-1}$  and  $\tilde{\mathbf{A}}_k = (\mathbf{0}_{m_n \times s_{nk}}, \mathbf{A})$  with  $s_{nk} = \|\hat{\mathbf{S}}_k\|_0$ . This implies that

$$T_{n,2} = 2^{-1} \left\| \sum_{k=1}^2 N^{-1/2} \mathbf{\Omega}_k^{1/2}(\boldsymbol{\theta}^*) \{N^{-1/2} \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + N^{1/2} \boldsymbol{\Delta}_n\} \right\|_2^2 + o_P(m_n),$$

which, combined with Lemma 11(iii), leads to the result of Corollary 2.  $\square$

**Proof of Theorem 5.** A specialization of Lemma 9 and  $\boldsymbol{\Delta}_n = \mathbf{0}$ .  $\square$

**Proofs of Theorems 5 and 6.** Theorem 5 be a specialization of Theorems 6, so we only prove Theorem 6. Since the random index sets  $\hat{\mathbf{S}}^{(1)}$  and  $\hat{\mathbf{S}}^{(2)}$  depend on the sub-samples  $\{Y_j^{(2)}, \mathbf{w}_j^{(2)}, j = 1, \dots, N\}$  and  $\{Y_j^{(1)}, \mathbf{w}_j^{(1)}, j = 1, \dots, N\}$ , respectively, it is challenging to establish the limiting distribution of the refitted statistic  $T$ , where  $T$  is  $T_{n,L}$ ,  $T_{n,W}$ , or  $T_{n,S}$ , since the degrees of freedom are random. We surmount this difficulty by using the Bahadur representation in Theorem 4, approximating  $T$  with a quadratic form in Lemma 11, and drawing a parallel to the quadratic form of the oracle GLR test statistic.

Denoted by  $\tilde{\mathbf{A}}_k = (\mathbf{0}_{m_n \times s_{nk}}, \mathbf{A})$ ,  $\tilde{\mathbf{A}}_o = (\mathbf{0}_{m_n \times s_n}, \mathbf{A})$ ,  $\mathbf{\Omega}_k(\boldsymbol{\theta}^*) = \{\tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k^\top\}^{-1}$ , and  $\mathbf{\Omega}_o(\boldsymbol{\theta}^*) = \{\tilde{\mathbf{A}}_o \mathbf{I}_o^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_o^\top\}^{-1}$ . Let

$$\begin{aligned} \boldsymbol{\xi}_i^{(k)} &= \mathbf{\Omega}_k^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*), \\ \text{and } \boldsymbol{\xi}_{i,o}^{(k)} &= \mathbf{\Omega}_o^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_o \mathbf{I}_o^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\mathbf{T}}^{(k)}(\boldsymbol{\theta}^*), \end{aligned}$$

where  $\mathbf{w}_{i,\hat{\mathbf{T}}}^{(k)}$  and  $\mathbf{w}_{i,\mathbf{T}}^{(k)}$  are subvectors of  $\mathbf{w}_i^{(k)}$  with indexes in  $\hat{\mathbf{T}}$  and  $\mathbf{T}$ , respectively.

Step (i). From Lemma 11, we obtain that

$$\begin{aligned} T &= \frac{1}{2} \left\| \sum_{k=1}^2 \mathbf{\Omega}_k^{0.5}(\boldsymbol{\theta}^*) [N^{-0.5} \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + N^{0.5} \boldsymbol{\Delta}_n] \right\|_2 + o_P(m_n) \\ &= \frac{1}{2} \left\| \sum_{k=1}^2 N^{-0.5} \sum_{i=1}^N \{\boldsymbol{\xi}_i^{(k)} + (\mathbf{\Omega}_k(\boldsymbol{\theta}^*))^{0.5} \boldsymbol{\Delta}_n\} \right\|_2^2 + o_P(m_n), \end{aligned} \quad (\text{A.15})$$

where  $T$  is  $T_{nL}$ ,  $T_{nW}$  or  $T_{nS}$ . In the following, we will prove that

$$T = T_* + o_P(m_n), \quad (\text{A.16})$$

where  $T_* = 2^{-1} \left\| \sum_{k=1}^2 N^{-0.5} \sum_{i=1}^N \{\boldsymbol{\xi}_{i,o}^{(k)} + (\mathbf{\Omega}_k(\boldsymbol{\theta}^*))^{0.5} \boldsymbol{\Delta}_n\} \right\|_2^2$ .

Case (a). Assume that  $m_n$  is fixed. Given  $k$  and  $\hat{\mathbf{S}}_k$ ,  $\{\boldsymbol{\xi}_i^{(k)}\}'s$  be iid with 0 mean and variance  $E\boldsymbol{\xi}_i^{(k)} \{\boldsymbol{\xi}_i^{(k)}\}^\top = \mathbb{I}_{m_n}$ , and  $\{\boldsymbol{\xi}_{i,o}^{(k)}\}'s$  be iid with 0 mean and variance  $E\boldsymbol{\xi}_{i,o}^{(k)} \{\boldsymbol{\xi}_{i,o}^{(k)}\}^\top = \mathbb{I}_{m_n}$ . By condition  $(B_2)$  and Theorem C (Serfling, 1980, p.36), with probability going to one, we have

$$\frac{\sum_{i=1}^N \boldsymbol{\xi}_i^{(k)}}{\sqrt{2N \log(\log \sqrt{N})}} \rightarrow \mathbf{1} \quad \text{and} \quad \frac{\sum_{i=1}^N \boldsymbol{\xi}_{i,o}^{(k)}}{\sqrt{2N \log(\log \sqrt{N})}} \rightarrow \mathbf{1}.$$

Put  $\bar{\boldsymbol{\Delta}}_{n,k} = (N\mathbf{\Omega}_k(\boldsymbol{\theta}^*))^{0.5} \boldsymbol{\Delta}_n$ . Note that  $\frac{\|\boldsymbol{\Delta}_n\|_2}{\sqrt{m_n/n}} = O(1)$  and  $\lambda_{\max}(\mathbf{\Omega}_k(\boldsymbol{\theta}^*)) = O(1)$  in Lemma 7, therefore

$$\|\bar{\boldsymbol{\Delta}}_{n,k}\|_2 \leq N^{0.5} \lambda_{\max}\{(\mathbf{\Omega}_k(\boldsymbol{\theta}^*))^{0.5}\} \|\boldsymbol{\Delta}_n\|_2 = O(\sqrt{m_n}) = O(1). \quad (\text{A.17})$$

Hence, with probability going to 1,

$$R = \frac{\|\sum_{k=1}^2 \{\bar{\Delta}_{n,k} + N^{-0.5} \sum_{i=1}^N \xi_{i,o}^{(k)}\}\|_2^2 / \log(\log \sqrt{N})}{\|\sum_{k=1}^2 \{\bar{\Delta}_{n,k} + N^{-0.5} \sum_{i=1}^N \xi_i^{(k)}\}\|_2^2 / \log(\log \sqrt{N})} \rightarrow 1.$$

This, combined with (A.15) and  $E\|\bar{\Delta}_{n,k} + N^{-0.5} \sum_{i=1}^N \xi_i^{(k)}\|_2^2 = O(m_n)$  in Lemma 8(iv), yields that

$$\begin{aligned} T &= T_* - 2^{-1} \left\| \sum_{k=1}^2 \{\bar{\Delta}_{n,k} + N^{-0.5} \sum_{i=1}^N \xi_i^{(k)}\} \right\|_2^2 (R - 1) + o_P(m_n) \\ &= T_* + o_P(m_n), \end{aligned}$$

which establishes (A.16).

Case (b). Assume that  $m_n \rightarrow \infty$ . Let  $L_{n,1} = \sum_{k=1}^2 N^{-1} \sum_{i,j=1}^N [\{\xi_j^{(k)}\}^\top \xi_i^{(k)} - \{\xi_{j,o}^{(k)}\}^\top \xi_{i,o}^{(k)}]$ ,  $L_{n,2} = 2 \sum_{k,l=1}^2 N^{-0.5} \sum_{i=1}^N \bar{\Delta}_{n,l}^\top \{\xi_i^{(k)} - \xi_{i,o}^{(k)}\}$ , and  $L_{n,3} = 2 \sum_{i=1}^N \sum_{j=1}^N N^{-1} [\{\xi_j^{(1)}\}^\top \xi_i^{(2)} - \{\xi_{j,o}^{(1)}\}^\top \xi_{i,o}^{(2)}]$ .

Then it can be rewritten that

$$\begin{aligned} \left\| \sum_{k=1}^2 N^{-0.5} \sum_{i=1}^N \{\xi_i^{(k)} + (\Omega_k(\theta^*))^{0.5} \Delta_n\} \right\|_2^2 &= \left\| N^{-0.5} \sum_{k=1}^2 \sum_{i=1}^N \{\xi_{i,o}^{(k)} + (\Omega_k(\theta^*))^{0.5} \Delta_n\} \right\|_2^2 \\ &\quad + L_{n,1} + L_{n,2} + L_{n,3}, \end{aligned}$$

and (A.16) holds if  $L_{n,j} = o_P(m_n)$  for  $j = 1, 2, 3$ . In the following we will show that  $L_{n,j} = o_P(m_n)$  for each  $j$ .

(b<sub>1</sub>). Let  $\Gamma_{i,j}^{(k)} = \{\xi_j^{(k)}\}^\top \xi_i^{(k)} - \{\xi_{j,o}^{(k)}\}^\top \xi_{i,o}^{(k)}$ . Then

$$\begin{aligned} L_{n,1} &= \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N [\{\xi_j^{(k)}\}^\top \xi_i^{(k)} - \{\xi_{j,o}^{(k)}\}^\top \xi_{i,o}^{(k)}] \\ &= \frac{1}{N} \sum_{i=1}^N \Gamma_{i,i}^{(k)} + \frac{1}{N} \sum_{i=1}^N \sum_{j=1, j \neq i}^N \Gamma_{i,j}^{(k)} \equiv L_{n,11} + L_{n,12}. \end{aligned}$$

Notice that, for  $i \neq j$ ,

$$E\xi_i^{(k)} \{\xi_i^{(k)}\}^\top = \mathbb{I}_{m_n}, \quad E\xi_{i,o}^{(k)} \{\xi_{i,o}^{(k)}\}^\top = \mathbb{I}_{m_n}, \quad E\xi_i^{(k)} \{\xi_j^{(k)}\}^\top = \mathbf{0}, \quad \text{and} \quad E\xi_{i,o}^{(k)} \{\xi_{j,o}^{(k)}\}^\top = \mathbf{0}. \quad (\text{A.18})$$

Then  $E\Gamma_{i,j} = 0$  for any  $1 \leq i, j \leq N$ . It follows from  $E(\xi_i^{(k)} \{\xi_i^{(k)}\}^\top \mid \hat{\mathbf{S}}_k) = \mathbb{I}_{m_n}$ , and conditional independence of  $\xi_i^{(k)}$  and  $\xi_j^{(k)}$  (given  $\hat{\mathbf{S}}_k$  for  $i \neq j$ ) that

$$\begin{aligned} E|\{\xi_j^{(k)}\}^\top \xi_i^{(k)}|^2 &= \text{Trace}[E\xi_i^{(k)} \{\xi_i^{(k)}\}^\top \xi_j^{(k)} \{\xi_j^{(k)}\}^\top] \\ &= \text{Trace}[E\{E(\xi_i^{(k)} \{\xi_i^{(k)}\}^\top \xi_j^{(k)} \{\xi_j^{(k)}\}^\top \mid \hat{\mathbf{S}}_k)\}] \\ &= \text{Trace}[E\{E(\xi_i^{(k)} \{\xi_i^{(k)}\}^\top \mid \hat{\mathbf{S}}_k) \cdot E(\xi_j^{(k)} \{\xi_j^{(k)}\}^\top \mid \hat{\mathbf{S}}_k)\}] \\ &= m_n \end{aligned}$$

for  $i \neq j$ . Similarly,  $E|\{\boldsymbol{\xi}_{j,o}^{(k)}\}^\top \boldsymbol{\xi}_{i,o}^{(k)}|^2 = m_n$  for  $i \neq j$ .  $\forall i, j \in [N] = \{1, \dots, N\}$ ,  $i \neq j$ , It follows that

$$\begin{aligned} E|\Gamma_{i,j}^{(k)}|^2 &= E|\{\boldsymbol{\xi}_j^{(k)}\}^\top \boldsymbol{\xi}_i^{(k)}|^2 + E|\{\boldsymbol{\xi}_{j,o}^{(k)}\}^\top \boldsymbol{\xi}_{i,o}^{(k)}|^2 - 2E[\{\boldsymbol{\xi}_j^{(k)}\}^\top \boldsymbol{\xi}_i^{(k)} \cdot \{\boldsymbol{\xi}_{j,o}^{(k)}\}^\top \boldsymbol{\xi}_{i,o}^{(k)}] \\ &= 2m_n - 2\text{Trace}(E[\boldsymbol{\xi}_i^{(k)}\{\boldsymbol{\xi}_{i,o}^{(k)}\}^\top] \cdot E[\boldsymbol{\xi}_i^{(k)}\{\boldsymbol{\xi}_{i,o}^{(k)}\}^\top]). \end{aligned}$$

which, combined with  $E[\boldsymbol{\xi}_i^{(k)}\{\boldsymbol{\xi}_{i,o}^{(k)}\}^\top] = (\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*))^{-1/2} \boldsymbol{\Omega}_o^{1/2}(\boldsymbol{\theta}^*)$ , yields that

$$\begin{aligned} E|\Gamma_{i,j}^{(k)}|^2 &= 2m_n - 2\text{Trace}\{(\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*))^{-0.5} \boldsymbol{\Omega}_o^{0.5}(\boldsymbol{\theta}^*)(\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*))^{-0.5} \boldsymbol{\Omega}_o^{0.5}(\boldsymbol{\theta}^*)\} \\ &= 2m_n - 2\text{Trace}\{(\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*))^{-0.5} \boldsymbol{\Omega}_o(\boldsymbol{\theta}^*)(\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*))^{-0.5}\}, \end{aligned}$$

where  $\text{Trace}(\mathbf{C}^\top \mathbf{C}) = \text{Trace}(\mathbf{C}^2)$ , for any square matrix  $\mathbf{C}$ . Thus,  $\text{Var}(\Gamma_{i,j}^{(k)}) = O(m_n)$  for  $i \neq j$ . Applying Lemma A (Serfling, 1980, p.185), we obtain that

$$\frac{1}{N\sqrt{\text{Var}(\Gamma_{i,j}^{(k)})}} \sum_{i=1}^N \sum_{j=1(\neq i)}^N \Gamma_{i,j}^{(k)} = O_P(1).$$

It follows that  $L_{n,12} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1(\neq i)}^N \Gamma_{i,j}^{(k)} = O_P(\sqrt{m_n}) = o_P(m_n)$ , if  $m_n \rightarrow \infty$ . Further, since  $E\Gamma_{i,i}^{(k)} = 0$ , it is easy to show that  $L_{n,11} = \frac{1}{N} \sum_{i=1}^N \Gamma_{i,i}^{(k)} = o_P(m_n)$ . Thus,  $L_{n,1} = o_P(m_n)$ .

( $b_2$ ). Obviously,  $L_{n,2} = 0$  if  $\bar{\Delta}_n = 0$ . In the following, we consider  $\bar{\Delta}_n \neq 0$ . Observe that

$$N^{-0.5} \left| \sum_{i=1}^N \bar{\Delta}_{n,l}^\top \boldsymbol{\xi}_i^{(k)} \right| = N^{-0.5} \|\bar{\Delta}_{n,l}\|_2 \left| \sum_{i=1}^N \bar{\Delta}_{n,l}^\top \boldsymbol{\xi}_i^{(k)} / \|\bar{\Delta}_{n,l}\|_2 \right|.$$

Using Chebyshev's inequality, we establish that

$$\begin{aligned} P\left(N^{-0.5} \left| \sum_{i=1}^N \bar{\Delta}_{n,l}^\top \boldsymbol{\xi}_i^{(k)} / \|\bar{\Delta}_{n,l}\|_2 \right| > a_n\right) &\leq a_n^{-2} N^{-1} \sum_{i=1}^N \sum_{j=1}^N E[\bar{\Delta}_{n,l}^\top \boldsymbol{\xi}_i^{(k)} \{\boldsymbol{\xi}_j^{(k)}\}^\top \bar{\Delta}_{n,l} / \|\bar{\Delta}_{n,l}\|_2^2] \\ &\leq a_n^{-2} N^{-1} \sum_{i=1}^N \sum_{j=1}^N \sup_{\mathbf{a} \in R^{m_n}, \|\mathbf{a}\|_2=1} |\mathbf{a}^\top E[\boldsymbol{\xi}_i^{(k)} \{\boldsymbol{\xi}_j^{(k)}\}^\top] \mathbf{a}|, \end{aligned}$$

for any  $a_n \rightarrow +\infty$ . This, combined with (A.18), yields that

$$P\left(N^{-0.5} \left| \sum_{i=1}^N \bar{\Delta}_{n,l}^\top \boldsymbol{\xi}_i^{(k)} / \|\bar{\Delta}_{n,l}\|_2 \right| > a_n\right) \leq a_n^{-2} \rightarrow 0.$$

Hence,  $N^{-0.5} \left| \sum_{i=1}^N \bar{\Delta}_{n,l}^\top \boldsymbol{\xi}_i^{(k)} / \|\bar{\Delta}_{n,l}\|_2 \right| = O_P(1)$ , which together with (A.17) yields

$$N^{-0.5} \left| \sum_{i=1}^N \bar{\Delta}_{n,l}^\top \boldsymbol{\xi}_i^{(k)} \right| = O_P(\sqrt{m_n}).$$

Similarly, We prove that  $N^{-0.5} \left| \sum_{i=1}^N \bar{\Delta}_{n,l}^\top \boldsymbol{\xi}_{i,o}^{(k)} \right| = O_P(\sqrt{m_n})$ . Therefore,

$$L_{n,2} = 2N^{-0.5} \sum_{k,l=1}^2 \sum_{i=1}^N \bar{\Delta}_{n,l}^\top \{\boldsymbol{\xi}_i^{(k)} - \boldsymbol{\xi}_{i,o}^{(k)}\} = O_P(\sqrt{m_n}) = o_P(m_n).$$

(b<sub>3</sub>). Rewrite  $L_{n,3}$  as

$$L_{n,3} = 2N^{-1} \sum_{i,j=1}^N \{\boldsymbol{\xi}_j^{(1)}\}^\top \boldsymbol{\xi}_i^{(2)} - 2N^{-1} \sum_{i,j=1}^N \{\boldsymbol{\xi}_{j,o}^{(1)}\}^\top \boldsymbol{\xi}_{i,o}^{(2)} \equiv L_{n,3}^{(1)} - L_{n,3}^{(2)}.$$

Computing the mean and variance, we can show that  $L_{n,3}^{(2)} = o_P(m_n)$ , since  $\boldsymbol{\xi}_{j,o}^{(1)}$  and  $\boldsymbol{\xi}_{i,o}^{(2)}$  are independent and  $E\boldsymbol{\xi}_{j,o}^{(1)} = E\boldsymbol{\xi}_{i,o}^{(2)} = \mathbf{0}$ , where  $i, j = 1, \dots, N$ . Notice that  $\|N^{-0.5} \sum_{i=1}^N \boldsymbol{\xi}_i^{(2)}\|_2 = O_P(m_n^{0.5})$  and

$$L_{n,3}^{(1)} = 2\{N^{-0.5} \sum_{i=1}^N \boldsymbol{\xi}_i^{(2)}\}^\top \{N^{-0.5} \sum_{j=1}^N \boldsymbol{\xi}_j^{(1)}\}.$$

We can replace  $\bar{\Delta}_n$  with  $N^{-0.5} \sum_{i=1}^N \boldsymbol{\xi}_i^{(2)}$  in Part (b<sub>2</sub>) and show that  $L_{n,3}^{(1)} = o_P(m_n)$ . Thus,  $L_{n,3} = o_P(m_n)$ .

Step (ii). Let  $\boldsymbol{\xi}_{i,o} = \{\boldsymbol{\xi}_{i,o}^{(1)} + \boldsymbol{\xi}_{i,o}^{(2)}\}/\sqrt{2}$ . Note that  $E\boldsymbol{\xi}_{i,o}^{(1)} = E\boldsymbol{\xi}_{i,o}^{(2)} = \mathbf{0}$  and  $\boldsymbol{\xi}_{i,o}^{(1)}$ , and  $\boldsymbol{\xi}_{i,o}^{(2)}$  are independent. It follows from (A.18) that  $\text{Cov}(\boldsymbol{\xi}_{i,o}) = E\boldsymbol{\xi}_{i,o}\boldsymbol{\xi}_{i,o}^\top = \mathbb{I}_{m_n}$ . Combining Cauchy-Schwarz inequality and Jessen's inequality, we can get

$$\begin{aligned} \sum_{i=1}^N E(\|\boldsymbol{\xi}_{i,o}\|_2^3) &\leq \sum_{i=1}^N E\{(\|\boldsymbol{\xi}_{i,o}^{(1)}\|_2^2 + \|\boldsymbol{\xi}_{i,o}^{(2)}\|_2^2)^{1.5}\} \leq 2 \sum_{i=1}^N \sum_{k=1}^2 E\|\boldsymbol{\xi}_{i,o}^{(k)}\|_2^3 \\ &= 2 \sum_{i=1}^N \sum_{k=1}^2 E\|\boldsymbol{\Omega}_o^{0.5}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_o \mathbf{I}_o^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_T^{(k)}(\boldsymbol{\theta}^*)\|_2^3. \end{aligned}$$

This, combined with  $\|\boldsymbol{\Omega}_o^{0.5}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_o \{\mathbf{I}_o(\boldsymbol{\theta}^*)\}^{-0.5}\|_2^2 \leq 1$ , yields that

$$\begin{aligned} \sum_{i=1}^N E(\|\boldsymbol{\xi}_{i,o}\|_2^3) &\leq 2 \sum_{i=1}^N \sum_{k=1}^2 E\|\{\mathbf{I}_o(\boldsymbol{\theta}^*)\}^{-0.5} \mathbf{U}_T^{(k)}(\boldsymbol{\theta}^*)\|_2^3 \\ &= 2 \sum_{i=1}^n E\|\{\mathbf{I}_o(\boldsymbol{\theta}^*)\}^{-0.5} \mathbf{U}_T(\boldsymbol{\theta}^*)\|_2^3. \end{aligned}$$

It follows from condition (B<sub>2</sub>) that

$$m_n^{0.25} N^{-1.5} \sum_{i=1}^n E\|\{\mathbf{I}_o(\boldsymbol{\theta}^*)\}^{-0.5} \mathbf{U}_T(\boldsymbol{\theta}^*)\|_2^3 \rightarrow 0.$$

Thus, by Theorem 1 of Bentkus (2004), we get

$$\sup_{\psi \in \mathcal{R}^{m_n}} |P(\sum_{i=1}^N N^{-0.5} \boldsymbol{\xi}_{i,o} \in \psi) - P(\mathcal{Z} \in \psi)| \leq c_* m_n^{0.25} N^{-1.5} \sum_{i=1}^N E(\|\boldsymbol{\xi}_{i,o}\|_2^3) \rightarrow 0, \quad (\text{A.19})$$

with  $\mathcal{Z} \in R^{m_n}$  be a Gaussian random vector for which  $E(\mathcal{Z}) = \mathbf{0}$ ,  $\text{Cov}(\mathcal{Z}) = \mathbb{I}_{m_n}$ , and the supremum is taken for all convex subsets in  $R^{m_n}$ . Now, let  $\Psi$  in (A.19) be the special convex set

$$\psi(x) = \left\{ \mathcal{T} \in R^{m_n} : \|\mathcal{T} + 2^{-0.5} \sum_{k=1}^2 \bar{\Delta}_{n,k}\|_2^2 \leq x \right\}.$$

Then

$$\sup_x |P(\sum_{i=1}^N N^{-0.5} \xi_{i,o} \in \psi(x)) - P(\mathcal{Z} \in \psi(x))| \rightarrow 0. \quad (\text{A.20})$$

Notice that  $T_* = \|\sum_{i=1}^N N^{-0.5} \xi_{i,o} + 2^{-0.5} \sum_{k=1}^2 \bar{\Delta}_{n,k}\|_2^2$ ,  $P\{\sum_{i=1}^N N^{-0.5} \xi_{i,o} \in \psi(x)\} = P(T_* \leq x)$ , and  $P\{\mathcal{Z} \in \psi(x)\} = P\{\chi_{m_n}^2(\eta_{n,\mathbf{L}}^2) \leq x\}$ , where  $\eta_{n,\mathbf{L}}^2 = \|2^{-0.5} \sum_{k=1}^2 \bar{\Delta}_{n,k}\|_2^2$ . In particular, (A.20) coincides with

$$\sup_x |P(T_* \leq x) - P\{\chi_{m_n}^2(\eta_{n,\mathbf{L}}^2) \leq x\}| \rightarrow 0,$$

which, together with (A.16), implies for every  $t \geq 0$ ,

$$\begin{aligned} P\{\chi_{m_n}^2(\eta_{n,\mathbf{L}}^2) \leq x - m_n t\} &\leq P(T_* \leq x - m_n t) + o(1) \\ &\leq P(T \leq x) + o(1) \\ &\leq P(T_* \leq x + m_n t) + o(1) \leq P\{\chi_{m_n}^2(\eta_{n,\mathbf{L}}^2) \leq x + m_n t\} + o(1). \end{aligned}$$

By Lemma S.7 in Shi et al. (2019), we have

$$\lim_{t \downarrow 0} |P(\chi_{m_n}^2(\eta_{n,\mathbf{L}}^2) \leq x - m_n t) - P(\chi_{m_n}^2(\eta_{n,\mathbf{L}}^2) \leq x + m_n t)| \rightarrow 0.$$

Therefore,  $\sup_x |P(T \leq x) - P(\chi_{m_n}^2(\eta_{n,\mathbf{L}}^2) \leq x)| \rightarrow 0$ . Notice that  $\bar{\Delta}_{n,k} = (N \Omega_k(\theta^*))^{0.5} \Delta_n$  with  $N = 0.5n$ , therefore

$$\eta_{n,\mathbf{L}}^2 = (4\phi)^{-1} n \left\| \sum_{k=1}^2 \Omega_k^{0.5}(\theta^*) \Delta_n \right\|_2^2 = (4\phi)^{-1} n \left\| \sum_{k=1}^2 \{\mathbf{A} \mathbf{I}_k^{22}(\theta^*) \mathbf{A}^\top\}^{-0.5} \Delta_n \right\|_2^2. \quad \square$$

**Proofs of Theorems 2 and 3.** Theorem 2 is a special case of Theorems 3, so we omit the argument for Theorem 2. According to Lemma 11(iv), we can get

$$T_{n,2} = \left\| \{\tilde{\mathbf{A}}_* \mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\theta^*) \tilde{\mathbf{A}}_*^\top\}^{-0.5} [n^{-0.5} \tilde{\mathbf{A}}_* \mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\theta^*) \mathbf{U}_{\hat{\mathbf{T}}}(\theta^*) + n^{0.5} \Delta_n] \right\|_2^2 + o_P(m_n), \quad (\text{A.21})$$

where  $\tilde{\mathbf{A}}_* = (\mathbf{0}_{m_n \times \|\hat{\mathbf{s}}\|_0}, \mathbf{A})$ . Note that, by Theorem 1,  $P(\hat{\mathbf{S}} = \mathbf{S}) \rightarrow 1$ , if condition (A<sub>6</sub>) holds. It follows that, with probability going to one,

$$T_{n,2} = \left\| \{\tilde{\mathbf{A}}_o \{\mathbf{I}_{\mathbf{T}}(\theta^*)\}^{-1} \tilde{\mathbf{A}}_o^\top\}^{-0.5} [n^{-0.5} \tilde{\mathbf{A}}_o \{\mathbf{I}_{\mathbf{T}}(\theta^*)\}^{-1} \mathbf{U}_{\hat{\mathbf{T}}}(\theta^*) + n^{0.5} \Delta_n] \right\|_2^2 + o_P(m_n),$$

where  $\tilde{\mathbf{A}}_o = (\mathbf{0}_{m_n \times s_n}, \mathbf{A})$ . Then, using the argument in step (ii) for the argument for Theorem 6, we can get  $\sup_x |P(T_{n,2} \leq x) - P(\chi_{m_n}^2(\eta_{n,o}^2) \leq x)| \rightarrow 0$ . This finishes the argument for the second result. For the first result, it can be proven similarly.  $\square$

**Proof of Theorems 7.** Recall that

$$\eta_{n,\mathbf{L}}^2 = (4\phi)^{-1}n \left\| \sum_{k=1}^2 \{\mathbf{A}\mathbf{I}_k^{22}(\boldsymbol{\theta}^*)\mathbf{A}^\top\}^{-0.5} \boldsymbol{\Delta}_n \right\|_2^2 \text{ and } \eta_{n,o}^2 = n\boldsymbol{\Delta}_n^\top \{\mathbf{A}\mathbf{I}_o^{22}(\boldsymbol{\theta}^*)\mathbf{A}^\top\}^{-1} \boldsymbol{\Delta}_n.$$

Then applying the Jensen inequality and Lemma 13(i), we obtain that

$$\eta_n^2 \leq \eta_{n,\mathbf{L}}^2 \leq 2^{-1} \sum_{k=1}^2 n\boldsymbol{\Delta}_n^\top \{\mathbf{A}\mathbf{I}_k^{22}(\boldsymbol{\theta}^*)\mathbf{A}^\top\}^{-1} \boldsymbol{\Delta}_n \leq n\boldsymbol{\Delta}_n^\top \{\mathbf{A}\mathbf{I}_o^{22}(\boldsymbol{\theta}^*)\mathbf{A}^\top\}^{-1} \boldsymbol{\Delta}_n = \eta_{n,o}^2.$$

Using the triangle inequality and Lemma 13(ii), we establish that

$$\sqrt{\eta_{n,o}^2} - \sqrt{\eta_{n,\mathbf{L}}^2} \leq (4\phi)^{-0.5} \sqrt{n} \left\| 2\{\mathbf{A}\mathbf{I}_o^{22}(\boldsymbol{\theta}^*)\mathbf{A}^\top\}^{-0.5} \boldsymbol{\Delta}_n - \sum_{k=1}^2 \{\mathbf{A}\mathbf{I}_k^{22}(\boldsymbol{\theta}^*)\mathbf{A}^\top\}^{-0.5} \boldsymbol{\Delta}_n \right\|_2 \leq \sqrt{4c^2 \eta_{n,\mathbf{L}}^2}.$$

Thus,  $\eta_{n,\mathbf{L}}^2 \geq (1 + 2c)^{-2} \eta_{n,o}^2 = K(\rho^*) \eta_{n,o}^2$ .  $\square$

## References

- [1] ANTONIADIS, A. AND FAN, J. (2001) Regularized wavelet approximations (with discussion). *Jour. Amer. Statist. Assoc.* **96**, 939-967.
- [2] BARTLETT, M. S. (1936). Some notes on insecticide tests in the laboratory and in the field. *Supplement to the Journal of the Royal Statistical Society.* **3**, 185-194.
- [3] BRADIC, J., FAN, J. AND JIANG, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39**, 3092-3120.
- [4] BRADIC, J., FAN, J., AND WANG, W. (2011). Penalized composite quasiLikelihood for ultrahigh-dimensional variable selection. *J. Roy. Statist. Soc. Ser. B.* **73**, 325-349.

SUPPLEMENTAL MATERIAL TO “DIMENSION-REDUCED TESTS FOR GENERALIZED QUASI-LIKELIHOOD MODELS BASED ON REGULARIZATION”

**Lemma 1.** Assume that (A.26)–(A.27) holds. Then

- (i) if (A.22) holds, conditions (A<sub>3</sub>) and (A<sub>4</sub>) hold for the linear regression;
- (ii) if  $\max_{1 \leq i \leq n} |\mathbf{w}_i^\top \boldsymbol{\theta}^*| = O(1)$  and (A.22) hold, conditions (A<sub>3</sub>) and (A<sub>4</sub>) hold for the logistic and Poisson regressions;
- (iii) if  $\max_{1 \leq i \leq n} |\mathbf{w}_i^\top \boldsymbol{\theta}^*| = O(1)$  and condition (A<sub>2</sub>) and (A.31) hold, conditions (A<sub>3</sub>) and (A<sub>4</sub>) hold for the Gamma and inverse Gaussian regression models with the log-link function.

**Proof of Lemma 1.** Firstly, we verify that, for  $m = O(s_n + q_n)$ ,

$$\inf_{\boldsymbol{\theta} \in \Pi_2} \min_{\mathbf{a} \in \Pi_1, \mathbf{a} \neq \mathbf{0}} \frac{\mathbf{a}^\top \mathbf{W}^\top \mathbf{W} \mathbf{a}}{n \|\mathbf{a}_T\|_2^2} \geq c, \quad \sup_{\boldsymbol{\theta} \in \Pi_2} \max_{\mathbf{a} \neq \mathbf{0}} \frac{\mathbf{a}^\top \mathbf{W}^\top \mathbf{W} \mathbf{a}}{n(\|\mathbf{a}_T\|_2^2 + \|\mathbf{a}_{T^c}\|_1^2)} = O(1), \quad (\text{A.22})$$

$$\psi_1(m) \equiv \max_{\|\mathbf{b}\|_0 \leq m, \mathbf{b} \neq \mathbf{0}} \frac{\mathbf{b}^\top \mathbf{W}_{T^c}^\top \mathbf{W}_{T^c} \mathbf{b}}{n \|\mathbf{b}\|_2^2} = O(s_n + q_n), \quad (\text{A.23})$$

$$\psi_2(m) \equiv \sup_{\boldsymbol{\theta} \in \Pi_2} \max_{j \in T} \max_{\|\mathbf{b}\|_0 \leq m, \mathbf{b} \neq \mathbf{0}} \frac{\mathbf{b}^\top \mathbf{W}_{T^c}^\top \text{diag}\{|\mathbf{W}_j|\} \mathbf{W}_{T^c} \mathbf{b}}{n \|\mathbf{b}\|_2^2} = O(\sqrt{n/c_n}), \quad (\text{A.24})$$

$$\psi_3(m) \equiv \sup_{\boldsymbol{\theta} \in \Pi_2} \max_{j \in T^c} \max_{\|\mathbf{b}\|_0 \leq m, \mathbf{b} \neq \mathbf{0}} \frac{\mathbf{b}^\top \mathbf{W}_{T^c}^\top \text{diag}\{|\mathbf{W}_j|\} \mathbf{W}_{T^c} \mathbf{b}}{n \|\mathbf{b}\|_2^2} = O\{(s_n + q_n)\sqrt{n/t_n}\}, \quad (\text{A.25})$$

where  $c_n \rightarrow \infty$  and  $c_n/\log n \rightarrow 0$ .

Applying Cauchy–Bunyakovsky–Schwarz, CBS inequality, we can get

$$\begin{aligned} N^{-1} \mathbf{a}^\top \mathbf{W}^\top \mathbf{W} \mathbf{a} &\leq 2N^{-1} (\mathbf{a}_T^\top \mathbf{W}_T^\top \mathbf{W}_T \mathbf{a}_T + \mathbf{a}_{T^c}^\top \mathbf{W}_{T^c}^\top \mathbf{W}_{T^c} \mathbf{a}_{T^c}) \\ &\leq 2\|N^{-1} \mathbf{W}_T^\top \mathbf{W}_T\|_2 \|\mathbf{a}_T\|_2^2 + 2N^{-1} \sum_{j,k \in T^c} |a_j a_k| \cdot |\mathbf{W}_j^\top \mathbf{W}_k| \\ &\leq 2\|N^{-1} \mathbf{W}_T^\top \mathbf{W}_T\|_2 \|\mathbf{a}_T\|_2^2 + 2\|\mathbf{a}_{T^c}\|_1^2 \max_{j \in T^c} N^{-1} \|\mathbf{W}_j\|_2^2, \end{aligned}$$

and

$$\begin{aligned} N^{-1} \mathbf{a}^\top \mathbf{W}^\top \mathbf{W} \mathbf{a} &\geq N^{-1} (\mathbf{a}_T^\top \mathbf{W}_T^\top \mathbf{W}_T \mathbf{a}_T + 2\mathbf{a}_T^\top \mathbf{W}_T^\top \mathbf{W}_{T^c} \mathbf{a}_{T^c}) \\ &\geq \lambda_{\min}(N^{-1} \mathbf{W}_T^\top \mathbf{W}_T) \|\mathbf{a}_T\|_2^2 - 2N^{-1} \sqrt{\mathbf{a}_T^\top \mathbf{W}_T^\top \mathbf{W}_T \mathbf{a}_T \cdot \mathbf{a}_{T^c}^\top \mathbf{W}_{T^c}^\top \mathbf{W}_{T^c} \mathbf{a}_{T^c}} \\ &\geq \lambda_{\min}(N^{-1} \mathbf{W}_T^\top \mathbf{W}_T) \|\mathbf{a}_T\|_2^2 - 2\|\mathbf{a}_T\|_2 \|\mathbf{a}_{T^c}\|_1 \max_{j \in T^c} \|N^{-1} \mathbf{W}_j\|_2 \|\mathbf{W}_T^\top \mathbf{W}_T\|_2^{1/2} \\ &= \lambda_{\min}(N^{-1} \mathbf{W}_T^\top \mathbf{W}_T) \|\mathbf{a}_T\|_2^2 - o(\|\mathbf{a}_T\|_2^2) \max_{j \in T^c} \|N^{-1} \mathbf{W}_j\|_2 \|\mathbf{W}_T^\top \mathbf{W}_T\|_2^{1/2}, \end{aligned}$$

where the last equality follows from  $\mathbf{a} \in \Pi_1$ . This, combined with  $\|N^{-1}\mathbf{W}_T^\top \mathbf{W}_T\|_2 = O(1)$ ,  $\lambda_{\min}(N^{-1}\mathbf{W}_T^\top \mathbf{W}_T) \geq c$ , and  $\max_{j \in T^c} N^{-1}\|\mathbf{W}_j\|_2^2 = O(1)$ , establishes (A.22). Note that

$$\begin{aligned}\psi_1(m) &\leq \max_{j \in T^c} N^{-1}\|\mathbf{W}_j\|_2^2 \times \sup_{\|\mathbf{b}\|_0 \leq m, \mathbf{b} \neq 0} \|\mathbf{b}\|_1^2 / \|\mathbf{b}\|_2^2 \leq m \max_{j \in T^c} N^{-1}\|\mathbf{W}_j\|_2^2, \\ \psi_2(m) &\leq \max_{j \in T^c} N^{-1}\|\mathbf{W}_j\|_2^2 \max_{j \in T} \|\mathbf{W}_j\|_\infty \times \sup_{\|\mathbf{b}\|_0 \leq m, \mathbf{b} \neq 0} \|\mathbf{b}\|_1^2 / \|\mathbf{b}\|_2^2 \\ &\leq m \max_{j \in T^c} N^{-1}\|\mathbf{W}_j\|_2^2 \max_{j \in T} \|\mathbf{W}_j\|_\infty, \\ \psi_3(m) &\leq \max_{j \in T^c} N^{-1}\|\mathbf{W}_j\|_2^2 \max_{j \in T^c} \|\mathbf{W}_j\|_\infty \times \sup_{\|\mathbf{b}\|_0 \leq m, \mathbf{b} \neq 0} \|\mathbf{b}\|_1^2 / \|\mathbf{b}\|_2^2 \\ &\leq m \max_{j \in T^c} N^{-1}\|\mathbf{W}_j\|_2^2 \max_{j \in T^c} \|\mathbf{W}_j\|_\infty.\end{aligned}$$

It follows from  $\frac{1}{\sqrt{n}} \max_{1 \leq j \leq p_n + q_n} \|\mathbf{W}_j\|_2 = O(1)$ ,  $c_n = O(\log n)$ ,

$$\max_{j \in T^c} \|\mathbf{W}_j\|_\infty = O(\sqrt{n/\log(r_n)}), \quad \max_{j \in T} \|\mathbf{W}_j\|_\infty = O(n^{1/2}(s_n + q_n)^{-1}(\log n)^{-1/2}), \quad (\text{A.26})$$

and  $m = O(s_n + q_n)$  that (A.23)-(A.25) hold.

- (i). For linear regression models, conditions (A<sub>3</sub>) and (A<sub>4</sub>) are implied by (A.22)-(A.25).
- (ii). For logistic regression,  $v_2(g^{-1}(t), y) = -\frac{1}{e^t + 2 + e^{-t}}$  and  $v_3(g^{-1}(t), y) = -\frac{e^t - e^{2t}}{1 + 3e^t + 3e^{2t} + e^{3t}}$ ; for Poisson regression,  $v_2(g^{-1}(t), y) = -\exp(t)$  and  $v_3(g^{-1}(t), y) = -\exp(t)$ . By (A.26),  $\boldsymbol{\theta} \in \Pi_2$  and  $c_n \leq \sqrt{\log n}$ , we have

$$\begin{aligned}\sup_{\boldsymbol{\theta} \in \Pi_2} |\mathbf{w}_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)| &\leq |\mathbf{w}_{i,T}^\top (\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^*)| + |\mathbf{w}_{i,T^c}^\top \boldsymbol{\theta}_{T^c}| \\ &\leq \sup_{\boldsymbol{\theta} \in \Pi_2} \|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^*\|_1 \max_{j \in T} \|\mathbf{W}_j\|_\infty + \sup_{\boldsymbol{\theta} \in \Pi_2} \|\boldsymbol{\theta}_{T^c}\|_1 \max_{j \in T^c} \|\mathbf{W}_j\|_\infty \\ &= (s_n + q_n)n^{-1/2} \sqrt{c_n} \max_{j \in T} \|\mathbf{W}_j\|_\infty + o(\sqrt{\log(r_n)/n})O(\sqrt{n/\log r_n}) \\ &= o(1),\end{aligned} \quad (\text{A.27})$$

uniformly for  $i = 1, \dots, n$  and  $\boldsymbol{\theta} \in \Pi_2$ , since  $\sup_{\boldsymbol{\theta} \in \Pi_2} \|\boldsymbol{\theta}_{T^c}\|_1 = \sup_{\boldsymbol{\theta} \in \Pi_2} \|\boldsymbol{\theta}_{T^c} - \boldsymbol{\theta}^*\|_1$ . This, combined with  $\max_{1 \leq i \leq n} |\mathbf{w}_i^\top \boldsymbol{\theta}^*| = O(1)$ , yields that

$$\sup_{\boldsymbol{\theta} \in \Pi_2} \max_{1 \leq i \leq n} |\mathbf{w}_i^\top \boldsymbol{\theta}| = O(1). \quad (\text{A.28})$$

Hence,  $\sup_{\boldsymbol{\theta} \in \Pi_2} \|\mathbf{v}_2(\boldsymbol{\theta})\|_2 = O(1)$ ,  $\inf_{\boldsymbol{\theta} \in \Pi_2} \min_{1 \leq i \leq n} |v_2(g^{-1}(\mathbf{w}_i^\top \boldsymbol{\theta}), Y_i)| > 0$ , and  $\sup_{\boldsymbol{\theta} \in \Pi_2} \|\mathbf{v}_3(\boldsymbol{\theta})\|_2 = O(1)$ . Then for logistic regression and Poisson regression models, conditions (A<sub>3</sub>) is implied by (A.22), and condition (A<sub>4</sub>) by (A.23)-(A.25).

- (iii). For the Gamma and inverse Gaussian regressions, the results can be proven similarly. In the following we only show the result for the Gamma regression.

(a) For Gamma regression models with the log-link function, we get

$$v_1(g^{-1}(t), y) = \frac{y - \exp(t)}{\exp(t)}, \quad v_2(g^{-1}(t), y) = -y \exp(-t), \quad \text{and} \quad v_3(g^{-1}(t), y) = y \exp(-t).$$

Then

$$\text{diag}\{\mathbf{v}_1(\boldsymbol{\theta}^*)\} = -\mathbf{v}_2(\boldsymbol{\theta}^*) - \mathbf{I}_n. \quad (\text{A.29})$$

Let  $\mathbf{u} = (u_1, \dots, u_n)^\top = \mathbf{W}\mathbf{a}$ . Then

$$\begin{aligned} |\mathbf{a}^\top \mathbf{W}^\top \{-\mathbf{v}_2(\boldsymbol{\theta}) + \mathbf{v}_2(\boldsymbol{\theta}^*)\} \mathbf{W}\mathbf{a}| &= |\mathbf{u}\{-\mathbf{v}_2(\boldsymbol{\theta}) + \mathbf{v}_2(\boldsymbol{\theta}^*)\}\mathbf{u}| \\ &= \left| \sum_{i=1}^n \{\exp(\mathbf{w}_i^\top \boldsymbol{\theta}^* - \mathbf{w}_i^\top \boldsymbol{\theta}) - 1\} u_i^2 Y_i \exp(-\mathbf{w}_i^\top \boldsymbol{\theta}^*) \right| \\ &\leq \max_{1 \leq i \leq n} |\exp(\mathbf{w}_i^\top \boldsymbol{\theta}^* - \mathbf{w}_i^\top \boldsymbol{\theta}) - 1| \cdot \left| \sum_{i=1}^n u_i^2 Y_i \exp(-\mathbf{w}_i^\top \boldsymbol{\theta}^*) \right| \end{aligned}$$

This, combined with (A.27), yields that

$$\sup_{\boldsymbol{\theta} \in \Pi_2} |\mathbf{a}^\top \mathbf{W}^\top \{-\mathbf{v}_2(\boldsymbol{\theta}) + \mathbf{v}_2(\boldsymbol{\theta}^*)\} \mathbf{W}\mathbf{a}| = o(1) \sup_{\boldsymbol{\theta} \in \Pi_2} \mathbf{a}^\top \mathbf{W}^\top \{-\mathbf{v}_2(\boldsymbol{\theta}^*)\} \mathbf{W}\mathbf{a}. \quad (\text{A.30})$$

Thus, condition (A<sub>3</sub>) holds if

$$\min_{\mathbf{a} \in \Pi_1, \mathbf{a} \neq \mathbf{0}} -\frac{\mathbf{a}^\top \mathbf{W}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}\mathbf{a}}{n\|\mathbf{a}_T\|_2^2} \geq \rho_1 \quad \text{and} \quad \max_{\mathbf{a} \neq \mathbf{0}} -\frac{\mathbf{a}^\top \mathbf{W}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}\mathbf{a}}{n(\|\mathbf{a}_T\|_2^2 + \|\mathbf{a}_{T^c}\|_1^2)} = O(1). \quad (\text{A.31})$$

It suffices to show that (A.31) holds. By (A.31), we have

$$\begin{aligned} -\mathbf{a}^\top \mathbf{W}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}\mathbf{a} &\geq -\mathbf{a}_T^\top \mathbf{W}_T^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}_T \mathbf{a}_T - 2\mathbf{a}_T^\top \mathbf{W}_T^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}_{T^c} \mathbf{a}_{T^c} \\ &\geq cn\|\mathbf{a}_T\|_2^2 - 2\sqrt{\mathbf{a}_T^\top \mathbf{W}_T^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}_T \mathbf{a}_T \cdot \mathbf{a}_{T^c}^\top \mathbf{W}_{T^c}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}_{T^c} \mathbf{a}_{T^c}} \\ &= cn\|\mathbf{a}_T\|_2^2 - O(\sqrt{n}\|\mathbf{a}_T\|_2) \sqrt{-\mathbf{a}_{T^c}^\top \mathbf{W}_{T^c}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}_{T^c} \mathbf{a}_{T^c}} \end{aligned} \quad (\text{A.32})$$

and

$$\begin{aligned} -\mathbf{a}^\top \mathbf{W}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}\mathbf{a} &\leq -2\mathbf{a}_T^\top \mathbf{W}_T^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}_T \mathbf{a}_T - 2\mathbf{a}_{T^c}^\top \mathbf{W}_{T^c}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}_{T^c} \mathbf{a}_{T^c} \\ &= O(n)\|\mathbf{a}_T\|_2 + \mathbf{a}_{T^c}^\top \mathbf{W}_{T^c}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}_{T^c} \mathbf{a}_{T^c}. \end{aligned} \quad (\text{A.33})$$

By (A.29), we get

$$\begin{aligned} -\mathbf{a}_{T^c}^\top \mathbf{W}_{T^c}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}_{T^c} \mathbf{a}_{T^c} &\leq \max_{j,k \in T^c} |\mathbf{W}_j^\top \{-\mathbf{v}_2(\boldsymbol{\theta}^*)\} \mathbf{W}_k| \cdot \|\mathbf{a}_{T^c}\|_1^2 \\ &= \max_{j,k \in T^c} |\mathbf{W}_j^\top \{-\mathbf{v}_2(\boldsymbol{\theta}^*) - \mathbf{I}_n\} \mathbf{W}_k + \mathbf{W}_j^\top \mathbf{W}_k| \cdot \|\mathbf{a}_{T^c}\|_1^2 \\ &\leq \max_{j,k \in T^c} \{|\mathbf{W}_j^\top \text{diag}\{\mathbf{v}_1(\boldsymbol{\theta}^*)\} \mathbf{W}_k| + \|\mathbf{W}_j\|_2^2\} \|\mathbf{a}_{T^c}\|_1^2. \end{aligned} \quad (\text{A.34})$$

Put  $c_{j,k} = \|\mathbf{W}_j \circ \mathbf{W}_k\|_2^2 v_0 + \|\mathbf{W}_j \circ \mathbf{W}_k\|_\infty n M_0$ , where  $v_0$  and  $M_0$  are defined in condition (A<sub>2</sub>).

It follows from (A.26) that

$$\begin{aligned} \max_{j,k \in T^c} c_{j,k} &\leq \max_{j,k \in T^c} \sum_{i=1}^n w_{i,j}^2 w_{i,k}^2 v_0 + \max_{j,k \in T^c} \max_{1 \leq i \leq n} |w_{i,j} w_{i,k}| n M_0 \\ &\leq \max_{j,k \in T^c} \|\mathbf{W}_j\|_2^2 \|\mathbf{W}_k\|_\infty^2 v_0 + \max_{j,k \in T^c} \|\mathbf{W}_j\|_\infty \|\mathbf{W}_k\|_\infty n M_0 \\ &= O(n^2 / \log r_n). \end{aligned}$$

Then using condition  $(A_2)$ , Bonferroni's inequality and (21) in Fan and Lv (2011), we obtain that, for any  $e_n \rightarrow \infty$ ,

$$\begin{aligned} P\left\{\max_{j,k \in \mathbf{T}^c} |\mathbf{W}_j^\top \text{diag}\{\mathbf{v}_1(\boldsymbol{\theta}^*)\} \mathbf{W}_k| > e_n n\right\} &\leq \sum_{j,k \in \mathbf{T}^c} P\left\{|\mathbf{W}_j^\top \text{diag}\{\mathbf{v}_1(\boldsymbol{\theta}^*)\} \mathbf{W}_k| > e_n n\right\} \\ &\leq 2 \sum_{j,k \in \mathbf{T}^c} \exp(-0.5 e_n n^2 / c_{j,k}) \\ &\leq 2 \exp\left(\log r_n - 0.5 e_n n^2 / \max_{j,k \in \mathbf{T}^c} c_{j,k}\right) \rightarrow 0. \end{aligned}$$

Hence,

$$\max_{j,k \in \mathbf{T}^c} |\mathbf{W}_j^\top \text{diag}\{\mathbf{v}_1(\boldsymbol{\theta}^*)\} \mathbf{W}_k| = O_P(n).$$

This, combined with (A.34),  $\max_{j,k \in \mathbf{T}^c} \|\mathbf{W}_j\|_2 = O(\sqrt{n})$  in condition  $(A_5)$ , yields that

$$-\mathbf{a}_{\mathbf{T}^c}^\top \mathbf{W}_{\mathbf{T}^c}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}_{\mathbf{T}^c} \mathbf{a}_{\mathbf{T}^c} = O(n) \|\mathbf{a}_{\mathbf{T}^c}\|_1^2, \quad (\text{A.35})$$

which, together with (A.33), leads to

$$\max_{\mathbf{a} \neq \mathbf{0}} -\frac{\mathbf{a}^\top \mathbf{W}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W} \mathbf{a}}{n(\|\mathbf{a}_{\mathbf{T}}\|_2^2 + \|\mathbf{a}_{\mathbf{T}^c}\|_1^2)} = O(1).$$

Applying (A.32) and  $\|\mathbf{a}_{\mathbf{T}^c}\|_1 = o(\|\mathbf{a}_{\mathbf{T}}\|_2^2)$  by  $\mathbf{a} \in \boldsymbol{\Pi}_1$ , we establish that

$$-\mathbf{a}^\top \mathbf{W}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W} \mathbf{a} \geq (c - o(1)) n \|\mathbf{a}_{\mathbf{T}}\|_2^2$$

for any  $\mathbf{a} \in \boldsymbol{\Pi}_1$ . Hence, There is a constant  $\rho_1 \in (0, c)$  for which

$$\min_{\mathbf{a} \in \boldsymbol{\Pi}_1, \mathbf{a} \neq \mathbf{0}} -\frac{\mathbf{a}^\top \mathbf{W}^\top \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W} \mathbf{a}}{n \|\mathbf{a}_{\mathbf{T}}\|_2^2} \geq \rho_1.$$

That is, (A.31) holds.

(b) Due to  $\text{diag}\{|\mathbf{v}_3(\boldsymbol{\theta})|\} = \mathbf{v}_2(\boldsymbol{\theta})$  for the Gamma regression, it holds that

$$\begin{aligned} \frac{\mathbf{b}^\top \mathbf{W}_{\mathbf{T}^c}^\top \text{diag}\{|\mathbf{v}_3(\boldsymbol{\theta})| \circ |\mathbf{W}_j|\} \mathbf{W}_{\mathbf{T}^c} \mathbf{b}}{n \|\mathbf{b}\|_2^2} &\leq \|\mathbf{W}_j\|_\infty \frac{\mathbf{b}^\top \mathbf{W}_{\mathbf{T}^c}^\top \text{diag}\{|\mathbf{v}_3(\boldsymbol{\theta})|\} \mathbf{W}_{\mathbf{T}^c} \mathbf{b}}{n \|\mathbf{b}\|_2^2} \\ &= \|\mathbf{W}_j\|_\infty \frac{-\mathbf{b}^\top \mathbf{W}_{\mathbf{T}^c}^\top \mathbf{v}_2(\boldsymbol{\theta}) \mathbf{W}_{\mathbf{T}^c} \mathbf{b}}{n \|\mathbf{b}\|_2^2}. \end{aligned}$$

By condition  $(A_3)$ , we have

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Pi}_2} \max_{\|\mathbf{b}\|_0 \leq m, \mathbf{b} \neq \mathbf{0}} \frac{-\mathbf{b}^\top \mathbf{W}_{\mathbf{T}^c}^\top \mathbf{v}_2(\boldsymbol{\theta}) \mathbf{W}_{\mathbf{T}^c} \mathbf{b}}{n \|\mathbf{b}\|_2^2} \leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Pi}_2} \max_{\|\mathbf{b}\|_0 \leq m, \mathbf{b} \neq \mathbf{0}} \frac{-m \mathbf{b}^\top \mathbf{W}_{\mathbf{T}^c}^\top \mathbf{v}_2(\boldsymbol{\theta}) \mathbf{W}_{\mathbf{T}^c} \mathbf{b}}{n \|\mathbf{b}\|_1^2} = O(m).$$

This, combined with (A.26), yields condition  $(A_4)$ .  $\square$

**Lemma 2.** Assume that  $\sup_{\|\mathbf{v}\|_2 \leq 1} \|\mathbf{v}^\top \mathbf{w}_i\|_{\psi_2} \leq \alpha_0$ ,  $\max_{1 \leq i \leq n} |\mathbf{w}_i^\top \boldsymbol{\theta}^*| = O(1)$ , and  $(s_n + q_n)^2 \{\log(nr_n)\}^{3/2} = o(n)$ , where  $\alpha_0$  is some positive constant and  $r_n = p_n + q_n$ . Then the following results hold:

- (a) condition  $(B_1)(iii)$  holds for linear, logistic and Poisson regression models;
- (b) if  $\max_{1 \leq j \leq r_n} \|w_{i,j} v_{1,i}(\boldsymbol{\theta}^*)\|_{\psi_1} \leq \alpha_1$  holds for some constant  $\alpha_1 > 0$ , then condition  $(B_1)(iii)$  holds for gamma and inverse Gaussian regression models.

**Proof of Lemma 2.** Note that  $(s_n + q_n)^2 \{\log(nr_n)\}^{3/2} = o(n)$ , it follows that

$$\sqrt{\log(n) \log(nr_n)}(s_n + q_n) = o(n^{1/2}) \quad \text{and} \quad (s_n + q_n) \log(nr_n) \log(r_n) = o(n). \quad (\text{A.36})$$

By (S5.41) in Shi et al. (2019), we have

$$P\left\{\max_{1 \leq i \leq n, 1 \leq j \leq r_n} |w_{i,j}| > c\sqrt{\log(nr_n)}\right\} \rightarrow 0 \quad (\text{A.37})$$

for any large constant  $c$ . Thus,

$$\sup_{\boldsymbol{\theta} \in \Pi_3} |\mathbf{w}_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)| \leq \sup_{\boldsymbol{\theta} \in \Pi_3} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 \times \max_{1 \leq i \leq n, 1 \leq j \leq r_n} |w_{i,j}| = O_P\{\sqrt{\log(nr_n)}\} \sup_{\boldsymbol{\theta} \in \Pi_3} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1.$$

This, combined with  $\max_{1 \leq i \leq n} |\mathbf{w}_i^\top \boldsymbol{\theta}^*| = O(1)$  and (A.36), yields that

$$\sup_{\boldsymbol{\theta} \in \Pi_3} |\mathbf{w}_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)| = o_P(1) \quad \text{and} \quad \sup_{\boldsymbol{\theta} \in \Pi_3} |\mathbf{w}_i^\top \boldsymbol{\theta}| = O_P(1). \quad (\text{A.38})$$

(a) Notice that  $v_3(g^{-1}(t), y) \equiv 0$  for linear regression models,  $v_3(g^{-1}(t), y) = -\frac{e^t - e^{2t}}{1 + 3e^t + 3e^{2t} + e^{3t}}$  for logistic regression models, and  $v_3(g^{-1}(t), y) = -\exp(t)$  for Poisson regression models. Since  $\sup_{\boldsymbol{\theta} \in \Pi_3} |\mathbf{w}_i^\top \boldsymbol{\theta}| = O_P(1)$ , condition  $(B_1)(iii)$  holds if

$$\sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}(N^{-1} \mathbf{W}_M^\top \text{diag}(|\mathbf{W}_j|) \mathbf{W}_M) = O_P(1), \quad (\text{A.39})$$

where  $\mathcal{Y} = \{\mathbf{M} \subseteq \{1, \dots, r_n\} : \mathbf{T} \subseteq \mathbf{M}, \|\mathbf{M}\|_0 \leq c(s_n + q_n)\}$  for some constant  $c > 0$ . Let  $t_1 = \sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \|E w_{i,j} \mathbf{w}_{i,\mathbf{M}} \mathbf{w}_{i,\mathbf{M}}^\top\|_2$  and

$$t_2 = \sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}\{N^{-1} \mathbf{W}_M^\top \text{diag}(|\mathbf{W}_j|) \mathbf{W}_M - E w_{i,j} \mathbf{w}_{i,\mathbf{M}} \mathbf{w}_{i,\mathbf{M}}^\top\}.$$

By norm subadditivity, we get

$$\sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}\{N^{-1} \mathbf{W}_M^\top \text{diag}(|\mathbf{W}_j|) \mathbf{W}_M\} \leq t_1 + t_2.$$

Using Cauchy–Schwarz,  $\sup_{\|\mathbf{v}\|_2 \leq 1} \|\mathbf{v}^\top \mathbf{w}_i\|_{\psi_2} \leq \alpha_0$ , and the Strum theorem about eigenvalues that

$$t_1 \leq \max_{1 \leq j \leq r_n} \|E w_{i,j} \mathbf{w}_i \mathbf{w}_i^\top\|_2 \leq \max_{1 \leq j \leq r_n} \sup_{\|\mathbf{v}\|_2=1} \sqrt{E|w_{i,j}|^2} \sqrt{E(\mathbf{v}^\top \mathbf{w}_i)^4} = O(1).$$

Let  $t_3 = \max_{1 \leq i \leq n, 1 \leq j \leq r_n} |w_{i,j}|$ . Using (A.37), we obtain that

$$\begin{aligned} P(t_2 \geq 2) &= P\{t_2 \geq 2, t_3 > c\sqrt{\log(nr_n)}\} + P\{t_2 \geq 2, t_3 \leq c\sqrt{\log(nr_n)}\} \\ &= P\{t_2 \geq 2, t_3 \leq c\sqrt{\log(nr_n)}\} + o(1). \end{aligned}$$

Then (A.39) holds if

$$P\left[\sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}\left\{N^{-1} \sum_{i=1}^n |w_{i,j}| \tilde{\mathbf{w}}_{i,\mathbf{M}} \tilde{\mathbf{w}}_{i,\mathbf{M}}^\top - E|w_{i,j}| \tilde{\mathbf{w}}_{i,\mathbf{M}} \tilde{\mathbf{w}}_{i,\mathbf{M}}^\top\right\} > 2\right] \rightarrow 0, \quad (\text{A.40})$$

where  $\tilde{\mathbf{w}}_{i,\mathbf{M}} = \mathbf{w}_{i,\mathbf{M}} \mathbf{1}(t_3 \leq c\sqrt{\log(nr_n)})$  and  $\mathbf{1}(t_3 \leq c\sqrt{\log(nr_n)})$  equals one when  $t_3 \leq c\sqrt{\log(nr_n)}$  and zero otherwise.

By the definition of the Orlicz norm and  $\sup_{\|\mathbf{v}\|_2 \leq 1} \|\mathbf{v}^\top \mathbf{w}_i\|_{\psi_2} \leq \alpha_0$ , we obtain that

$$\max_{1 \leq j \leq r_n} \|w_{i,j}\|_{\psi_1} \leq \max_{1 \leq j \leq r_n} \|w_{i,j}\|_{\psi_2} / \sqrt{\log 2} \leq \alpha_0 / \sqrt{\log 2}.$$

Then, following the argument for Lemma C.2 in Shi et al. (2018), For all  $j$  and  $\mathbf{M}$ , we can get

$$P\left\{\lambda_{\max}\left(N^{-1} \sum_{i=1}^n |w_{i,j}| \tilde{\mathbf{w}}_{i,\mathbf{M}} \tilde{\mathbf{w}}_{i,\mathbf{M}}^\top - E|w_{i,j}| \tilde{\mathbf{w}}_{i,\mathbf{M}} \tilde{\mathbf{w}}_{i,\mathbf{M}}^\top\right) > 2\right\} \leq \|\mathbf{M}\|_0 \exp\left\{-\frac{2n^2}{\tilde{\alpha}_0 \|\mathbf{M}\|_0 n \sqrt{\log(nr_n)}}\right\}.$$

where  $\tilde{\alpha}_0 = 2c\alpha_0/\sqrt{\log 2} + 4c^3\alpha_0^3/\log 2$ . This, combined with Bonferroni's inequality and (A.36), yields that

$$\begin{aligned} &P\left[\sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}\left\{N^{-1} \sum_{i=1}^n |w_{i,j}| \tilde{\mathbf{w}}_{i,\mathbf{M}} \tilde{\mathbf{w}}_{i,\mathbf{M}}^\top - E|w_{i,j}| \tilde{\mathbf{w}}_{i,\mathbf{M}} \tilde{\mathbf{w}}_{i,\mathbf{M}}^\top\right\} > 2\right] \\ &\leq \sum_{\mathbf{M} \in \mathcal{Y}, 1 \leq j \leq r_n} P\left[\lambda_{\max}\left\{N^{-1} \sum_{i=1}^n |w_{i,j}| \tilde{\mathbf{w}}_{i,\mathbf{M}} \tilde{\mathbf{w}}_{i,\mathbf{M}}^\top - E|w_{i,j}| \tilde{\mathbf{w}}_{i,\mathbf{M}} \tilde{\mathbf{w}}_{i,\mathbf{M}}^\top\right\} > 2\right] \\ &= cr_n^{c(s_n+q_n)+1} (s_n + q_n) \exp\left\{-\frac{2n^2}{\tilde{\alpha}_0 c(s_n + q_n) n \sqrt{\log(nr_n)}}\right\} \rightarrow 0. \end{aligned}$$

This establishes (A.40).

(b) We only prove the result for the gamma regression. Similarly, we can prove the result for the inverse Gaussian regression. For the gamma regression,  $v_3(g^{-1}(t), y) = y \exp(-t)$ . Similar to (A.30), we can show that

$$\|\mathbf{W}_\mathbf{M}^\top \{\text{diag}(|\mathbf{v}_3(\boldsymbol{\theta})| \circ |\mathbf{W}_j|) - \text{diag}(|\mathbf{v}_3(\boldsymbol{\theta}^*)| \circ |\mathbf{W}_j|)\} \mathbf{W}_\mathbf{M}\|_2 = o_P(1) \|\mathbf{W}_\mathbf{M}^\top \text{diag}(|\mathbf{v}_3(\boldsymbol{\theta}^*)| \circ |\mathbf{W}_j|) \mathbf{W}_\mathbf{M}\|_2,$$

uniformly for  $\mathbf{M} \in \mathcal{Y}$ . Then  $(B_1)(iii)$  follows if

$$\sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}\{\mathbf{W}_{\mathbf{M}}^{\top} \text{diag}(|\mathbf{v}_3(\boldsymbol{\theta}^*)| \circ |\mathbf{W}_j|) \mathbf{W}_{\mathbf{M}}\} = O_P(n). \quad (\text{A.41})$$

Since  $v_1(g^{-1}(t), y) = y \exp(-t) - 1$ , it holds that

$$\begin{aligned} & \sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}\{\mathbf{W}_{\mathbf{M}}^{\top} \text{diag}(|\mathbf{v}_3(\boldsymbol{\theta}^*)| \circ |\mathbf{W}_j|) \mathbf{W}_{\mathbf{M}}\} \\ & \leq \sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}\{\mathbf{W}_{\mathbf{M}}^{\top} \text{diag}(|\mathbf{W}_j|) \mathbf{W}_{\mathbf{M}}\} \\ & \quad + \sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}\left[\mathbf{W}_{\mathbf{M}}^{\top} \{\text{diag}(|\mathbf{v}_3(\boldsymbol{\theta}^*)| \circ |\mathbf{W}_j|) - \text{diag}(|\mathbf{W}_j|)\} \mathbf{W}_{\mathbf{M}}\right] \\ & = \sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}\{\mathbf{W}_{\mathbf{M}}^{\top} \text{diag}(|\mathbf{W}_j|) \mathbf{W}_{\mathbf{M}}\} + \sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}\left[\mathbf{W}_{\mathbf{M}}^{\top} \text{diag}\{\mathbf{v}_1(\boldsymbol{\theta}^*) \circ |\mathbf{W}_j|\} \mathbf{W}_{\mathbf{M}}\right]. \end{aligned}$$

Since  $\max_{1 \leq j \leq r_n} \|w_{i,j} v_{1,i}(\boldsymbol{\theta}^*)\|_{\psi_1} \leq \alpha_1$ , similar to (A.39), we get

$$\sup_{\mathbf{M} \in \mathcal{Y}} \max_{1 \leq j \leq r_n} \lambda_{\max}\{\mathbf{W}_{\mathbf{M}}^{\top} \text{diag}(\mathbf{v}_1(\boldsymbol{\theta}^*) \circ |\mathbf{W}_j|) \mathbf{W}_{\mathbf{M}}\} = O_P(n).$$

Thus, condition  $(B_2)(iii)$  holds for Gamma regression models.  $\square$

**Lemma 3.** Given any  $k \in \mathbb{N}$  and  $\ell \geq 1$ , we have

$$\frac{\varphi(\lceil lk \rceil)}{\varphi(k)} \leq \lceil l \rceil,$$

where  $\varphi(\cdot)$  is  $\varphi_1(\cdot)$ ,  $\varphi_2(\cdot)$  or  $\varphi_3(\cdot)$ , and  $\lceil l \rceil = \min\{a | a \geq l, a \in \mathbb{N}_+\}$  with  $\mathbb{N}_+$  being the positive integer set.

**Proof of Lemma 3.** Since  $-\mathbf{W}_{\mathbf{T}^c}^{\top} \mathbf{v}_2(\boldsymbol{\theta}^*) \mathbf{W}_{\mathbf{T}^c}$  and  $\mathbf{W}_{\mathbf{T}^c}^{\top} \text{diag}\{|\mathbf{v}_3(\boldsymbol{\theta}^*)| \circ |\mathbf{W}_j|\} \mathbf{W}_{\mathbf{T}^c}$  are positive semidefinite matrices, the result can be proven along the same line for the proof of Lemma 3 in Belloni and Chernozhukov (2013).  $\square$

**Lemma 4.** Let  $\boldsymbol{\delta} = \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$  and  $t_n = \min\{\log(r_n), s_n + q_n\}$ , where  $r_n = p_n + q_n$ . Under conditions  $(A_1)$  to  $(A_5)$ . We can get

- (i)  $\|\mathbf{U}(\boldsymbol{\theta}^*)\|_{\infty} = O_P\{\sqrt{\log(r_n)/n}\}$  and  $\|U_{n,\mathbf{T}}(\boldsymbol{\theta}^*)\|_2 = O_P\{\sqrt{(s_n + q_n)/n}\}$ ;
- (ii)  $\|\boldsymbol{\delta}_{\mathbf{T}}\|_2 = O_P\{\sqrt{(s_n + q_n)/n}\}$  and  $\|\boldsymbol{\delta}_{\mathbf{T}^c}\|_1 = O_P\{t_n/(n\lambda_n)\}$ .

**Proof of Lemma 4.** (i) Under conditions  $(A_2)$  and  $(A_5)$ , using (S5.13) in the supplementary of Shi et al. (2019), we establish that\*\*\*

$$\|\mathbf{U}(\boldsymbol{\theta}^*)\|_{\infty} = O_P(\sqrt{\log(q_n + p_n)/n}). \quad (\text{A.42})$$

Applying Chebyshev's inequality and condition (A<sub>2</sub>), we obtain that

$$\begin{aligned} P\{\|\mathbf{U}_T(\boldsymbol{\theta}^*)\|_2^2 \geq a_n n(s_n + q_n)\} &\leq \frac{\text{Trace}[E\{\mathbf{U}_T(\boldsymbol{\theta}^*)\}^{\otimes 2}]}{a_n n(s_n + q_n)} \\ &\leq \frac{O(1)(s_n + q_n)}{a_n(s_n + q_n)} \rightarrow 0 \end{aligned}$$

for any  $a_n \rightarrow \infty$ . Hence,

$$\|\mathbf{U}_T(\boldsymbol{\theta}^*)\|_2 = O_P(\sqrt{(s_n + q_n)/n}). \quad (\text{A.43})$$

(ii) Define

$$\boldsymbol{\Omega}_{\tau_n} = \{\boldsymbol{\theta}^* + \mathbf{a} : \|\mathbf{a}_T\|_2 \leq \tau_n \sqrt{\frac{s_n + q_n}{n}}, \|\mathbf{a}_{T^c}\|_1 \leq \frac{\tau_n t_n}{n \lambda_n}, \text{ and } \mathbf{a} \in \mathbf{R}^{p_n + q_n}\},$$

where  $\tau_n \rightarrow \infty$ ,  $\tau_n \sqrt{(s_n + q_n)/n} \leq d_n$ ,  $\frac{\tau_n t_n}{n \lambda_n} = o(\sqrt{n^{-1/2} \log(p_n + q_n)})$ , and  $\tau_n / \sqrt{\log n} \rightarrow 0$ . By the concavity of  $Q_n(\boldsymbol{\theta})$ , It is enough to verify that, w.h.p., a local maximizer  $\tilde{\boldsymbol{\theta}}$  lies in the interior of  $\boldsymbol{\Omega}_{\tau_n}$ , or equivalently

$$Q_n(\boldsymbol{\theta}^*) - n \sum_{j=1}^{p_n} p_{\lambda_n}(|\theta_j^*|) - \{Q_n(\boldsymbol{\theta}^* + \mathbf{a}) - n \sum_{j=1}^{p_n} p_{\lambda_n}(|\theta_j^* + a_j|)\} > 0, \quad (\text{A.44})$$

for all  $\boldsymbol{\theta}^* + \mathbf{a} \in \partial \boldsymbol{\Omega}_{\tau_n}$ , where  $\partial \boldsymbol{\Omega}_{\tau_n}$  is the boundary of closed set  $\boldsymbol{\Omega}_{\tau_n}$ . Recall that  $\mathbf{v}_1(\boldsymbol{\theta}) = (v_1(\mathbf{w}_1^\top \boldsymbol{\theta}, Y_1), \dots, v_1(\mathbf{w}_n^\top \boldsymbol{\theta}, Y_n))^\top$ ,  $\mathbf{v}_2(\boldsymbol{\theta}) = \text{diag}(v_2(\mathbf{w}_1^\top \boldsymbol{\theta}, Y_1), \dots, v_2(\mathbf{w}_n^\top \boldsymbol{\theta}, Y_n))$  with  $v_1(t) = \frac{\partial Q(t, y)}{\partial t}$  and  $v_2(t) = \frac{\partial^2 Q(t, y)}{\partial t^2}$ . By Taylor's expansion, we can get

$$Q_n(\boldsymbol{\theta}^* + \mathbf{a}) - Q_n(\boldsymbol{\theta}^*) = \mathbf{a}^\top \mathbf{U}(\boldsymbol{\theta}^*) + 0.5 \mathbf{a}^\top \mathbf{H}(\boldsymbol{\theta}_0) \mathbf{a}, \quad (\text{A.45})$$

where  $\boldsymbol{\theta}_0^* \in [\boldsymbol{\theta}^*, \boldsymbol{\theta}^* + \mathbf{a}]$ . Because  $\lambda_n^{-1} \sqrt{\log(q_n + p_n)/n} \rightarrow 0$ , by (A.42), for some  $c > 0$ ,

$$\lambda_n \geq c \|n^{-1} \mathbf{U}(\boldsymbol{\theta}^*)\|_\infty. \quad (\text{A.46})$$

Combining Hölder's inequality with (A.46), we can get

$$\begin{aligned} \mathbf{a}^\top \mathbf{U}(\boldsymbol{\theta}^*) &\leq \mathbf{a}_T^\top \mathbf{U}_T(\boldsymbol{\theta}^*) + |\mathbf{a}_{T^c}^\top \mathbf{U}_{T^c}(\boldsymbol{\theta}^*)| \\ &\leq \mathbf{a}_T^\top \mathbf{U}_T(\boldsymbol{\theta}^*) + n \|N^{-1} \mathbf{U}(\boldsymbol{\theta}^*)\|_\infty \|\mathbf{a}_{T^c}\|_1 \\ &\leq \mathbf{a}_T^\top \mathbf{U}_T(\boldsymbol{\theta}^*) + c^{-1} n \lambda_n \|\mathbf{a}_{T^c}\|_1. \end{aligned} \quad (\text{A.47})$$

Combining (A.45) and (A.47) leads to

$$Q_n(\boldsymbol{\theta}^*) - Q_n(\boldsymbol{\theta}^* + \mathbf{a}) \geq -0.5 \mathbf{a}^\top \mathbf{H}(\boldsymbol{\theta}_0^*) \mathbf{a} - \mathbf{a}_T^\top \mathbf{U}_T(\boldsymbol{\theta}^*) - c^{-1} n \lambda_n \|\mathbf{a}_{T^c}\|_1,$$

which, together with condition (A<sub>3</sub>) and the definition of  $\partial\Omega_{\tau_n}$ , yields that

$$\begin{aligned} Q_n(\boldsymbol{\theta}^*) - Q_n(\boldsymbol{\theta}^* + \mathbf{a}) &\geq 0.5n\rho_1\|\mathbf{a}_T\|_2^2 - c^{-1}n\lambda_n\|\mathbf{a}_{T^c}\|_1 - \mathbf{a}_T^\top \mathbf{U}_T(\boldsymbol{\theta}^*) \\ &\geq 0.5\rho_1\tau_n^2(s_n + q_n) - c^{-1}\tau_n(s_n + q_n) - \mathbf{a}_T^\top \mathbf{U}_T(\boldsymbol{\theta}^*). \end{aligned} \quad (\text{A.48})$$

Applying the mean-value formula, we have

$$\begin{aligned} \sum_{i=1}^{p_n} p_{\lambda_n}(|\theta_i^*|) - \sum_{i=1}^{p_n} p_{\lambda_n}(|\theta_i^* + a_j|) &= - \sum_{i=1}^{p_n} p'_{\lambda_n}(|\check{\theta}_i|) a_i \text{sgn}(\check{\theta}_i) \\ &= - \sum_{i \in \mathbf{S}} p'_{\lambda_n}(|\check{\theta}_i|) a_i \text{sgn}(\check{\theta}_i) - \lambda_n \sum_{i \in \mathbf{T}^c} \rho'(|\check{\theta}_i|, \lambda_n) a_i \text{sgn}(\check{\theta}_i) \\ &= -\mathbf{a}_S^\top \bar{\rho}_{\lambda_n}(|\check{\boldsymbol{\theta}}_S|) - \lambda_n \sum_{i \in \mathbf{T}^c} \rho'(|\check{\theta}_i|, \lambda_n) a_i \text{sgn}(\check{\theta}_i), \end{aligned} \quad (\text{A.49})$$

with  $\check{\boldsymbol{\theta}} \in [\boldsymbol{\theta}^*, \boldsymbol{\theta}^* + \mathbf{a}]$  and  $\bar{\rho}_{\lambda_n}(|\check{\boldsymbol{\theta}}_S|) = (p'_{\lambda_n}(|\check{\theta}_i|) \text{sgn}(\check{\theta}_i), i \in \mathbf{S})^\top$ . By condition (A<sub>3</sub>), we have

$$\|\check{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^*\|_\infty \leq \|\mathbf{a}_S\|_\infty \leq \|\mathbf{a}_S\|_2 = \tau_n \sqrt{(s_n + q_n)/n} \leq d_n,$$

which implies that

$$|\check{\theta}_i| \geq |\theta_i^*| - d_n \geq \min_{i \in \mathbf{S}} |\theta_i^*| - d_n = d_n. \quad (\text{A.50})$$

Since  $\rho'_{\lambda_n}(t) \geq 0$  and  $\rho''_{\lambda_n}(t) \leq 0$  for  $t \in [0, \infty)$  for given  $\lambda_n$ , which implies that

$$|\sum_{i \in \mathbf{T}^c} \rho'_{\lambda_n}(|\check{\theta}_i|) a_i \text{sgn}(\check{\theta}_i)| \leq \sum_{i \in \mathbf{T}^c} \rho'_{\lambda_n}(|\check{\theta}_i|) |a_i| \leq \sum_{i \in \mathbf{T}^c} \rho'(0+) |a_i| = \rho'(0+) \|\mathbf{a}_{T^c}\|_1,$$

which together with (A.49) and the definition of  $\Omega_{\tau_n}$ , yields that

$$\begin{aligned} n \sum_{i=1}^{p_n} p_{\lambda_n}(|\theta_i^* + a_i|) - n \sum_{i=1}^{p_n} p_{\lambda_n}(|\theta_i^*|) &= n\mathbf{a}_S^\top \bar{\rho}_{\lambda_n}(|\check{\boldsymbol{\theta}}_S|) + n\lambda_n \sum_{i \in \mathbf{T}^c} \rho'_{\lambda_n}(|\check{\theta}_i|) a_i \text{sgn}(\check{\theta}_i) \\ &\geq n\mathbf{a}_S^\top \bar{\rho}_{\lambda_n}(|\check{\boldsymbol{\theta}}_S|) - n\lambda_n \rho'(0+) \|\mathbf{a}_{T^c}\|_1 \\ &\geq n\mathbf{a}_S^\top \bar{\rho}_{\lambda_n}(|\check{\boldsymbol{\theta}}_S|) - \rho'(0+) \tau_n t_n. \end{aligned} \quad (\text{A.51})$$

Under (A.48) and (A.51), we obtain that

$$\begin{aligned} Q_n(\boldsymbol{\theta}^*) - Q_n(\boldsymbol{\theta}^* + \mathbf{a}) + n \sum_{i=1}^{p_n} p_{\lambda_n}(|\theta_i^* + a_i|) - n \sum_{i=1}^{p_n} p_{\lambda_n}(|\theta_i^*|) \\ \geq 0.5\rho_1\tau_n^2(s_n + q_n) - c^{-1}\tau_n(s_n + q_n) - \mathbf{a}_T^\top \mathbf{U}_{n,T}(\boldsymbol{\theta}^*) + n\mathbf{a}_S^\top \bar{\rho}_{\lambda_n}(|\check{\boldsymbol{\theta}}_S|) - \rho'(0+)\tau_n(s_n + q_n) \\ = 0.5\rho_1\tau_n^2(s_n + q_n) - (c^{-1} + \rho'(0+))\tau_n(s_n + q_n) - \mathbf{a}_T^\top \mathbf{U}_T(\boldsymbol{\theta}^*) + n\mathbf{a}_S^\top \bar{\rho}_{\lambda_n}(|\check{\boldsymbol{\theta}}_S|). \end{aligned} \quad (\text{A.52})$$

Note that  $p'_{\lambda_n}(t)$  is decreasing in  $t \in [0, \infty)$  and  $p'_{\lambda_n}(d_n) = O(n^{-1/2})$ , it follows from Hölder's inequality, (A.50), and the definition of  $\mathbf{\Omega}_{\tau_n}$  that

$$\begin{aligned} \|\mathbf{n}\mathbf{a}_S^\top \bar{\rho}_{\lambda_n}(|\check{\boldsymbol{\theta}}_S|)\|_2^2 &\leq \|\mathbf{a}_S\|_2^2 n^2 \sum_{i \in S} \{p'_{\lambda_n}(|\check{\theta}_i|)\}^2 \\ &\leq n\tau_n^2(s_n + q_n) \sum_{i \in S} \{p'_{\lambda_n}(d_n)\}^2 \\ &= O\{\tau_n^2(s_n + q_n)^2\}. \end{aligned} \quad (\text{A.53})$$

Again, from Hölder's inequality and  $\mathbf{\Omega}_{\tau_n}$ , yields

$$\|\mathbf{a}_T^\top \mathbf{U}_T(\boldsymbol{\theta}^*)\|_2 \leq \|\mathbf{a}_T\|_2 \|\mathbf{U}_T(\boldsymbol{\theta}^*)\|_2 = \tau_n \sqrt{N^{-1}(s_n + q_n)} \|\mathbf{U}_T(\boldsymbol{\theta}^*)\|_2 = O_P\{\tau_n(s_n + q_n)\}.$$

This, combined with (A.52) and (A.53), establishes (A.44).  $\square$

**Lemma 5.** Under conditions  $(A_1)$ – $(A_5)$  and  $(B_1)$ . For both  $k = 1$  and  $k = 2$ ,

$$\|N^{-1}\mathbf{H}_{\mathbf{T}_k}^{(k)}(\boldsymbol{\theta}, t) + \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 = O_P\{n^{-0.5}(s_n + q_n)\}.$$

**Proof of Lemma 5.** By condition  $(B_1)$ , therefore

$$\begin{aligned} &E(\|N^{-1}\mathbf{H}_{\mathbf{T}_k}^{(k)}(\boldsymbol{\theta}, t) + \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2^2) \\ &\leq N^{-2} \sum_{j, l \in \mathbf{S}_k \cup \mathbf{D}} E \left[ \sum_{i=1}^N \left\{ w_{i,j}^{(k)} w_{i,l}^{(k)} v_2(\{\mathbf{w}_i^{(k)}\}^\top \boldsymbol{\theta}^*, Y_i^{(k)}) - E[w_{i,j}^{(k)} w_{i,l}^{(k)} v_2(\{\mathbf{w}_i^{(k)}\}^\top \boldsymbol{\theta}^*, Y_i^{(k)})] \right\}^2 \right] \\ &= O\{N^{-1}(s_{nk} + q_n)^2\}, \end{aligned} \quad (\text{A.54})$$

Since the summands are independent (hence uncorrelated),  $\text{Var}(\sum_i X_i) = \sum_i \text{Var}(X_i)$ . By Theorem 1, we know that  $s_{nk} + q_n = O(s_n + q_n)$  with probability going to 1. This, combined with (A.54), completes the proof.  $\square$

**Lemma 6.** Under conditions  $(A_1)$  to  $(A_5)$  and  $(B_1)$ , If  $\frac{s_n + q_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), and there are two constant numbers  $C_1, C_2 \in (0, \infty)$ , independent of  $n$ , for which  $\|\mathbf{A}\mathbf{A}^\top\|_{\text{op}} \leq C_1$  and  $\|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_{\text{op}} \leq C_2$ , and  $\|\boldsymbol{\Delta}_n\|_2 = O(\sqrt{m_n/n})$ , then for  $k = 1, 2$ ,

- (i) under  $H_{a,n}^{(2)}$ ,  $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\|_2 = O_P(\sqrt{(s_n + q_n)/n})$ ;
- (ii) under  $H_{0,2}$ ,  $\|\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*\|_2 = O_P(\sqrt{(s_n + q_n)/n})$ .

**Proof of Lemma 6.** (i) Define

$$\mathbb{H} = \{\boldsymbol{\theta}^* + \mathbf{a} : \|\mathbf{a}\|_2 = \tau_n \sqrt{(s_n + q_n)/n}, \mathbf{a}_{\hat{\mathbf{T}}_k^c} = \mathbf{0}\},$$

where  $\tau_n \rightarrow \infty$ ,  $\tau_n/\sqrt{\log n} \rightarrow 0$ , and  $\tau_n(s_n + q_n)/n \rightarrow 0$ . By the concavity of  $Q_n^{(k)}(\boldsymbol{\theta})$ , It will be enough to verify that, w.p.a.1, the objective attains a local optimum at some  $\hat{\boldsymbol{\theta}} \in \text{int}(\mathbb{H})$ . That is,

$$P\{\max_{\boldsymbol{\theta} \in \mathbb{H}} Q_n^{(k)}(\boldsymbol{\theta}) - Q_n^{(k)}(\boldsymbol{\theta}^*) < 0\} \rightarrow 1. \quad (\text{A.55})$$

By the mean value theorem and the CBS inequality, we can get

$$\begin{aligned} Q_n^{(k)}(\boldsymbol{\theta}) - Q_n^{(k)}(\boldsymbol{\theta}^*) &= U_{n, \hat{\mathbf{T}}_k}^{(k)} \mathbf{a}_{\hat{\mathbf{T}}_k} + \frac{1}{2} \mathbf{a}_{\hat{\mathbf{T}}_k}^\top \mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}_*) \mathbf{a}_{\hat{\mathbf{T}}_k} \\ &\leq \tau_n \sqrt{(s_n + q_n)/n} \|\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}\|_2 + \frac{1}{2} \mathbf{a}_{\hat{\mathbf{T}}_k}^\top \mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}_*) \mathbf{a}_{\hat{\mathbf{T}}_k} \\ &\equiv \tau_n \sqrt{(s_n + q_n)/n} I_4 - I_5, \end{aligned}$$

with  $\boldsymbol{\theta}_* \in [\boldsymbol{\theta}^*, \boldsymbol{\theta}]$ . Hence, (A.55) holds if

$$P[\max_{\mathbf{a}} \{\tau_n \sqrt{(s_n + q_n)/n} I_4 - I_5\} < 0] \rightarrow 1, \text{ as } n \rightarrow \infty. \quad (\text{A.56})$$

By Theorem 1 and  $s_n + q_n = o(n^{1/3})$ , we get

$$s_{nk} + q_n = O(s_n + q_n) = o(n^{1/3}). \quad (\text{A.57})$$

Applying Taylor's expansion, condition  $(B_1)(iii)$ ,  $(\boldsymbol{\theta}_* - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k^c} = \mathbf{0}$ , and  $\|(\boldsymbol{\theta}_* - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k}\|_2 \leq \|\mathbf{a}_{\hat{\mathbf{T}}_k}\|_2 = O(\tau_n \sqrt{(s_n + q_n)/n})$ , we obtain that

$$\begin{aligned} &\|N^{-1}\{\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}_*) - \mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\}\|_2 \\ &\leq \max_{\|\mathbf{b}\|_2=1} \mathbf{b}^\top \mathbb{W}_k^\top \text{diag}\{|\mathbf{v}_{3,k}(\boldsymbol{\theta}_0)| \circ |\mathbb{W}_k(\boldsymbol{\theta}_* - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k}|\} \mathbb{W}_k \mathbf{b} / N \\ &\leq \|(\boldsymbol{\theta}_* - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k}\|_1 N^{-1} \max_{j \in \hat{\mathbf{T}}_k} \lambda_{\max}[\mathbf{W}_{\hat{\mathbf{T}}_k}^\top \text{diag}\{|\mathbf{v}_3(\boldsymbol{\theta}_0)| \circ |\mathbf{W}_j|\} \mathbf{W}_{\hat{\mathbf{T}}_k}] \\ &= O_P\{\tau_n(s_n + q_n)n^{-1/2}\}, \end{aligned} \quad (\text{A.58})$$

where  $\boldsymbol{\theta}_0$  is between  $\boldsymbol{\theta}_*$  and  $\boldsymbol{\theta}^*$ . This, combined with  $\tau_n(s_n + q_n)/n \rightarrow 0$  and Lemma 5, yields that

$$\|N^{-1}\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}_*) + \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 = O_P\{\tau_n(s_n + q_n)n^{-1/2}\} = o_P(1).$$

Note that  $\lambda_{\min}(\mathbf{I}_k(\boldsymbol{\theta}^*)) \geq 1/\rho_4$  in condition  $(B_1)$ , it follows that

$$\lambda_{\min}\left(-N^{-1}\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}_*)\right) \geq (2\rho_4)^{-1},$$

with probability going to one. For the term  $I_5$ , we have  $2N^{-1}I_5 \geq (2\rho_4)^{-1}\|\mathbf{a}\|_2^2 = (2\rho_4)^{-1}\tau_n^2(s_n + q_n)/n$ . This implies that

$$I_5 \geq (8\rho_4)^{-1}\tau_n^2(s_n + q_n).$$

Since  $E\{\mathbf{U}^{(k)}(\boldsymbol{\theta}^*)\}^{\otimes 2} = \phi \text{Trace}(N\mathbf{I}_k(\boldsymbol{\theta}^*))$ , by Chebyshev's inequality, we get

$$P(I_4^2 \geq \tau_n n(s_n + q_n)) \leq \{\tau_n n(s_n + q_n)\}^{-1} E(I_4^2) = \{\tau_n n(s_n + q_n)\}^{-1} \phi \text{Trace}(N\mathbf{I}_k(\boldsymbol{\theta}^*)).$$

By Condition  $(B_1)$  and (A.57), we have

$$\text{Trace}(\mathbf{I}_k(\boldsymbol{\theta}^*)) \leq (s_{nk} + q_n)\lambda_{\max}(\mathbf{I}_k(\boldsymbol{\theta}^*)) = O\{\rho_4(s_n + q_n)\}.$$

This, combined with  $\tau_n \rightarrow \infty$ , yields that

$$P\{I_4^2 \geq \tau_n n(s_{nk} + q_n)\} \leq \frac{O(\tau_n n(s_n + q_n))}{\tilde{\rho}^2 n(s_n + q_n)} \rightarrow 0.$$

Then  $|I_4| < \sqrt{\tau_n n(s_n + q_n)}$  with probability going to 1. Therefore,  $\tau_n \sqrt{(s_n + q_n)/n} I_4 = o_P(I_5)$ . Since  $I_5 > 0$ , (A.56) holds.

(ii) Similar to Lemma 6 (i), the result holds.  $\square$

**Lemma 7.** Let  $\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}) = \{\tilde{\mathbf{A}}_k(\hat{\mathbf{I}}_k(\boldsymbol{\theta}))^{-1} \tilde{\mathbf{A}}_k^\top\}^{-1}$ . Under conditions  $(A_1)$ – $(A_5)$  and  $(B_1)$ . If  $\frac{s_n + q_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), and there are two constant numbers  $C_1, C_2 \in (0, \infty)$ , independent of  $n$ , for which  $\|\mathbf{A}\mathbf{A}^\top\|_{\text{op}} \leq C_1$  and  $\|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_{\text{op}} \leq C_2$ , then the following results hold:

- (i)  $\|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 = O_P((s_n + q_n)n^{-0.5})$ ,  $\lambda_{\max}(\mathbf{I}_k(\boldsymbol{\theta}^*)) = O(1)$ ,  $\lambda_{\max}(\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)) = O_P(1)$ ,  $\lambda_{\max}(\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)) = O(1)$ , and  $\lambda_{\max}(\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)) = O_P(1)$ ;
- (ii)  $\|\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) - \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\|_2 = O_P((s_n + q_n)n^{-0.5})$ ,  $\lambda_{\max}(\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)) = O(1)$ ,  $\lambda_{\max}(\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*)) = O_P(1)$ ,  $\|\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\|_2 = O_P(1)$ ,  $\|\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\|_2 = O(1)$ ,  $\|\tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\|_2 = O_P(1)$ , and  $\|\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\|_2 = O(1)$ ;
- (iii)  $\|\tilde{\mathbf{A}}_k^\top (\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) - \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*))\|_2 = O_P\{(s_n + q_n)n^{-0.5}\}$ , and  $\|\tilde{\mathbf{A}}_k^\top (\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) - \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*))\tilde{\mathbf{A}}_k\|_2 = O_P\{(s_n + q_n)n^{-0.5}\}$ .

**Proof of Lemma 7.** (i) From condition  $(B_1)$  and (A.57), we obtain that

$$1/\rho_4 \leq \lambda_{\min}(\mathbf{I}_k(\boldsymbol{\theta}^*)) \leq \lambda_{\max}(\mathbf{I}_k(\boldsymbol{\theta}^*)) \leq \rho_4.$$

In view of 5,  $\|\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*) - \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 = O_P(n^{-0.5}(s_n + q_n))$ , which, combined with  $s_n + q_n = o(n^{1/3})$ , leads to

$$\begin{aligned}\lambda_{\max}(\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)) &\leq \lambda_{\max}(\mathbf{I}_k(\boldsymbol{\theta}^*)) + \|\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*) - \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 = O_P(1); \\ \lambda_{\min}(\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)) &\geq \lambda_{\min}(\mathbf{I}_k(\boldsymbol{\theta}^*)) - \|\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*) - \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 \geq 1/\rho_1 - o_P(1).\end{aligned}$$

Thus,  $\lambda_{\min}(\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)) = O_P(1)$  and  $\lambda_{\max}(\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)) = O_P(1)$ . This, combined with the CBS inequality, yields

$$\|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \leq \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*) - \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 = O_P(n^{-0.5}(s_n + q_n)).$$

(ii) Notice that

$$\begin{aligned}\{\lambda_{\max}(\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*))\}^{-1} &= \inf_{\|\mathbf{a}\|_2=1} \mathbf{a}^\top \tilde{\mathbf{A}}_k(\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*))^{-1} \tilde{\mathbf{A}}_k^\top \mathbf{a} \\ &\geq \lambda_{\min}\{(\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*))^{-1}\} \inf_{\|\mathbf{a}\|_2=1} \mathbf{a}^\top \tilde{\mathbf{A}}_k \tilde{\mathbf{A}}_k^\top \mathbf{a} \\ &= \lambda_{\min}(\tilde{\mathbf{A}}_k \tilde{\mathbf{A}}_k^\top) / \lambda_{\max}(\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)).\end{aligned}$$

It follows from  $\lambda_{\max}((\mathbf{A}\mathbf{A}^\top)^{-1}) = O(1)$  and result (i) that

$$\lambda_{\max}(\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*)) \leq \lambda_{\max}((\mathbf{A}\mathbf{A}^\top)^{-1}) \lambda_{\max}(\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)) = O_P(1).$$

Similarly, we get that  $\lambda_{\max}(\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)) = O(1)$ . Using the CBS inequality and

$$\|\hat{\boldsymbol{\Omega}}_k^{0.5}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{-0.5}\|_2 = \|\hat{\boldsymbol{\Omega}}_k^{0.5}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{-1} \tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k^{0.5}(\boldsymbol{\theta}^*)\|_2^{0.5} = 1,$$

we establish that

$$\begin{aligned}\|\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 &= \|\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{-0.5} \{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{0.5}\|_2 \leq \|\{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{0.5}\|_2 \cdot \|(\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*))^{0.5}\|_2; \\ \|\tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 &= \|\{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{0.5} \{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{-0.5} \tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{-0.5} \{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{0.5}\|_2 \\ &\leq \|\{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{0.5}\|_2^2,\end{aligned}$$

where the last inequality holds because  $\{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{-0.5} \tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \{\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{-0.5}$  is a projection matrix. Thus,  $\|\tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 = O_P(1)$  and  $\|\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 = O_P(1)$ . Similarly, we can prove that  $\|\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 = O(1)$ , and  $\|\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 = O(1)$ .

With  $\|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 = O_P((s_n + q_n)n^{-0.5})$ ,  $\|\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 = O_P(1)$  and  $\|\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 = O(1)$ , we establish that

$$\begin{aligned}\|\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) - \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\|_2 &= \|\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k [\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)] \tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\|_2 \quad (\text{A.59}) \\ &\leq \|\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 \cdot \|\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 \cdot \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \\ &= O_P((s_n + q_n)n^{-0.5}).\end{aligned}$$

(iii) Combining the Hölder inequality with (A.59), we can get

$$\|\tilde{\mathbf{A}}_k^\top (\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) - \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)) \tilde{\mathbf{A}}_k\|_2 \leq \|\tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 \cdot \|\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 \cdot \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2,$$

and

$$\|\tilde{\mathbf{A}}_k^\top (\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) - \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*))\|_2 \leq \|\tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k\|_2 \cdot \|\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\|_2 \cdot \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2.$$

This, combined with results (i)-(ii), establishes result (iii).  $\square$

**Lemma 8.** Under conditions  $(A_1)$ – $(A_5)$  and  $(B_1)$ . If  $\frac{s_n + q_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), and there are two constant numbers  $C_1, C_2 \in (0, \infty)$ , independent of  $n$ , for which  $\|\mathbf{A}\mathbf{A}^\top\|_{\text{op}} \leq C_1$  and  $\|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_{\text{op}} \leq C_2$ . Under  $H_{a,2}^{(n)}$  and  $\frac{\|\boldsymbol{\Delta}_n\|_2}{\sqrt{m_n/n}} = O(1)$ , we obtain

- (i)  $(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k^c} = \mathbf{0}$ , and  $(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} = \{N\mathbf{I}_k(\boldsymbol{\theta}^*)\}^{-1} \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{r}_{n,1}^{(k)}$ , where  $\|\mathbf{r}_{n,1}^{(k)}\|_2 = o_P(n^{-0.5})$ ;
- (ii)  $(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k^c} = \mathbf{0}$  and

$$\begin{aligned} (\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} &= N^{-1} \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \boldsymbol{\Delta}_n \\ &\quad - N^{-1} \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{r}_{n,2}^{(k)}, \end{aligned}$$

where  $\|\mathbf{r}_{n,2}^{(k)}\|_2 = o_P(n^{-1/2})$ ;

- (iii)  $\|(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{0,k})_{\hat{\mathbf{T}}_k}\|_2 = O_P(\sqrt{m_n/n})$  and

$$(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{0,k})_{\hat{\mathbf{T}}_k} = N^{-1} \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \boldsymbol{\Delta}_n + \mathbf{r}_{n,3}^{(k)},$$

where  $\mathbf{r}_{n,3}^{(k)} = \mathbf{r}_{n,1}^{(k)} - \mathbf{r}_{n,2}^{(k)}$ ;

- (iv)  $\|\{\mathbf{I}_k(\boldsymbol{\theta}^*)\}^{-1/2} \tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) [N^{-1/2} \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + N^{1/2} \boldsymbol{\Delta}_n]\|_2 = O_P(\sqrt{m_n})$ .

**Proof of Lemma 8.** (i)-(ii). Results (i) and (ii) can be proven similarly, so we only show result (ii). Notice that the optimization problem

$$\max_{\boldsymbol{\theta}} \{Q_n^{(k)}(\boldsymbol{\theta}), \text{ s.t. } \boldsymbol{\beta}_{\hat{\mathbf{S}}_k^c} = \mathbf{0}, \mathbf{A}(\boldsymbol{\gamma} - \mathbf{t}) = \mathbf{0}\}$$

is equivalent to

$$\min_{\boldsymbol{\theta}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2} \{-Q_n^{(k)}(\boldsymbol{\theta}) + \boldsymbol{\mu}_1^\top \mathbf{A}(\boldsymbol{\gamma} - \mathbf{t}) + \boldsymbol{\mu}_2^\top \boldsymbol{\beta}_{\hat{\mathbf{S}}_k^c} + 0.5f_1 \|\mathbf{A}(\boldsymbol{\gamma} - \mathbf{t})\|_2^2 + 0.5f_2 \|\boldsymbol{\beta}_{\hat{\mathbf{S}}_k^c}\|_2^2\},$$

where  $\boldsymbol{\mu}_1 \in R^{m_n}$ ,  $\boldsymbol{\mu}_2 \in R^{p_n - s_{nk}}$ , and  $f_1, f_2$  are some positive constants. The definition of  $\hat{\boldsymbol{\theta}}_{0,k}$  and the first order stationary condition imply that

(a)  $(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} = \mathbf{0}$ ,  $\mathbf{A}(\hat{\boldsymbol{\gamma}}_{0,k} - \mathbf{t}) = \mathbf{0}$ , and  $\partial Q_n^{(k)}(\hat{\boldsymbol{\theta}}_{0,k})/\partial \boldsymbol{\theta}_{\hat{\mathbf{S}}_k} = \mathbf{0}$ .

(b) There exists some  $\boldsymbol{\mu}_1 \in R^{m_n}$  satisfies that  $\partial Q_n^{(k)}(\hat{\boldsymbol{\theta}}_{0,k})/\partial \boldsymbol{\theta}_{\mathbf{D}} = \mathbf{A}^\top \boldsymbol{\mu}_1$ . That is,

$$U_{\hat{\mathbf{T}}_k}^{(k)}(\hat{\boldsymbol{\theta}}_{0,k}) = \tilde{\mathbf{A}}_k^\top \tilde{\boldsymbol{\mu}}_1, \quad (\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} = \mathbf{0}, \quad \text{and} \quad \mathbf{A}(\hat{\boldsymbol{\gamma}}_{0,k} - \mathbf{t}) = \mathbf{0}, \quad (\text{A.60})$$

where  $\tilde{\boldsymbol{\mu}}_1 = (\mathbf{0}^\top, \boldsymbol{\mu}_1^\top)^\top$ .

Applying the Taylor expansion yields that

$$U_{\hat{\mathbf{T}}_k}^{(k)}(\hat{\boldsymbol{\theta}}_{0,k}) = U_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} + 0.5\mathbf{R}_{n,k}, \quad (\text{A.61})$$

where  $\mathbf{R}_{n,k} = \{(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k}^\top \mathbb{W}_k^\top \text{diag}(\mathbf{v}_{3,k}(\tilde{\boldsymbol{\theta}}_k) \circ \mathbb{W}_{k,j}) \mathbb{W}_k(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k}, 1 \leq j \leq s_{nk} + q_n\}^\top$  with  $\tilde{\boldsymbol{\theta}}_k$  lying in the line segment joining  $\hat{\boldsymbol{\theta}}_{0,k}$  and  $\boldsymbol{\theta}^*$ . Hence,

$$(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} = \{N\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{-1} \{U_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) - \tilde{\mathbf{A}}_k^\top \tilde{\boldsymbol{\mu}}_1 + 0.5\mathbf{R}_{n,k}\}, \quad (\text{A.62})$$

where the identity  $N\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*) = -\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)$  is used. By (A.60) and  $\mathbf{A}(\boldsymbol{\gamma}^* - \mathbf{t}) = \boldsymbol{\Delta}_n$ , we have  $\tilde{\mathbf{A}}_k(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} = -\boldsymbol{\Delta}_n$ . Thus, solving (A.62) for  $\tilde{\boldsymbol{\mu}}_1$ , we get

$$\tilde{\boldsymbol{\mu}}_1 = N\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*)[\boldsymbol{\Delta}_n + \tilde{\mathbf{A}}_k\{N\hat{\mathbf{I}}_k(\boldsymbol{\theta}^*)\}^{-1}\{U_{\hat{\mathbf{T}}_k}^{(k)}(\hat{\boldsymbol{\theta}}_{0,k}) + 0.5\mathbf{R}_{n,k}\}]. \quad (\text{A.63})$$

Then replacing  $\tilde{\boldsymbol{\mu}}_1$  in (A.62) by (A.63) gives us that

$$\begin{aligned} (\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} &= N^{-1}\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)U_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) - \hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*)\boldsymbol{\Delta}_n \\ &\quad - N^{-1}\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)U_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) \\ &\quad + 0.5N^{-1}\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)\{\mathbf{R}_{n,k} - \tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)\mathbf{R}_{n,k}\} \\ &\equiv \sum_{j=1}^4 \mathbf{F}_{k,j}. \end{aligned}$$

Let  $\mathbf{G}_{k,1} = N^{-1}\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)U_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)$ ,  $\mathbf{G}_{k,2} = -\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\boldsymbol{\Delta}_n$ , and

$$\mathbf{G}_{k,3} = -N^{-1}\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)U_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*).$$

Then

$$(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} = \sum_{j=1}^3 \mathbf{G}_{k,j} + \mathbf{r}_{n,2}^{(k)},$$

where  $r_{n,2}^{(k)} = \sum_{j=1}^3 (\mathbf{F}_{k,j} - \mathbf{G}_{k,j}) + \mathbf{F}_{k,4}$ . From here, it is enough to establish that  $\|\mathbf{r}_{n,2}^{(k)}\|_2 = o_P(n^{-1/2})$ . In the following, we prove that  $n^{1/2}\|\mathbf{F}_{k,j} - \mathbf{G}_{k,j}\|_2 \xrightarrow{p} 0$  with  $j$  can be 1, 2, and 3, and  $\|\mathbf{F}_{k,4}\|_2 = o_P(n^{-1/2})$ , which leads to result (ii).

First, using condition  $(B_1)$  and (A.57), we get

$$\text{Trace}\{E(U_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*))^{\otimes 2}\} = \text{Trace}\{N\phi\mathbf{I}_k(\boldsymbol{\theta}^*)\} = O(n(s_n + q_n)). \quad (\text{A.64})$$

Thus,

$$\|\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2 = O_P(\sqrt{n(s_n + q_n)}). \quad (\text{A.65})$$

This, combined with Lemma 7(i) and  $s_n + q_n = o(n^{1/3})$ , yields that

$$\begin{aligned} \|\mathbf{F}_{k,1} - \mathbf{G}_{k,1}\|_2 &\leq \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|N^{-1}\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2 \\ &= O_P((s_n + q_n)n^{-0.5}) \cdot O_P(\sqrt{(s_n + q_n)/n}) = o_P(n^{-0.5}). \end{aligned}$$

Second, by Lemma 7 and  $\|\Delta_n\|_2 = O(\sqrt{m_n/n})$ , it holds that

$$\begin{aligned} \|\mathbf{F}_{k,2} - \mathbf{G}_{k,2}\|_2 &= \|[\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)]\tilde{\mathbf{A}}_k^\top \hat{\Omega}_k(\boldsymbol{\theta}^*)\Delta_n + \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top (\hat{\Omega}_k(\boldsymbol{\theta}^*) - \Omega_k(\boldsymbol{\theta}^*))\Delta_n\|_2 \\ &\leq \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|\tilde{\mathbf{A}}_k^\top \hat{\Omega}_k(\boldsymbol{\theta}^*)\|_2 \cdot \|\Delta_n\|_2 \\ &\quad + \|\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|\tilde{\mathbf{A}}_k^\top (\hat{\Omega}_k(\boldsymbol{\theta}^*) - \Omega_k(\boldsymbol{\theta}^*))\|_2 \cdot \|\Delta_n\|_2 \\ &= O_P((s_n + q_n)n^{-0.5}) \cdot O_P(\sqrt{m_n/n}). \end{aligned}$$

Since  $m_n \leq s_n + q_n = o(n^{1/3})$ , we have  $\|\mathbf{F}_{k,2} - \mathbf{G}_{k,2}\|_2 = o_P(n^{-0.5})$ .

Thirdly, using the subadditivity of the norm, we establish that

$$\begin{aligned} \|\mathbf{F}_{k,3} - \mathbf{G}_{k,3}\|_2 &\leq \|N^{-1}[\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)]\tilde{\mathbf{A}}_k^\top \hat{\Omega}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2 \\ &\quad + \|N^{-1}\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top (\hat{\Omega}_k(\boldsymbol{\theta}^*) - \Omega_k(\boldsymbol{\theta}^*))\tilde{\mathbf{A}}_k\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2 \\ &\quad + \|N^{-1}\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top \Omega_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k[\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)]\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2 \\ &\leq \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|\tilde{\mathbf{A}}_k^\top \hat{\Omega}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\|_2 \cdot \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|N^{-1}\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2 \\ &\quad + \|\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|\tilde{\mathbf{A}}_k^\top (\hat{\Omega}_k(\boldsymbol{\theta}^*) - \Omega_k(\boldsymbol{\theta}^*))\tilde{\mathbf{A}}_k\|_2 \cdot \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|N^{-1}\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2 \\ &\quad + \|\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|\tilde{\mathbf{A}}_k^\top \Omega_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\|_2 \cdot \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|N^{-1}\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2. \end{aligned}$$

This, combined with Lemma 7 and (A.65), yields that

$$\|\mathbf{F}_{k,3} - \mathbf{G}_{k,3}\|_2 = O_P((s_n + q_n)n^{-0.5}) \cdot O_P(\sqrt{(s_n + q_n)/n}) = o_P(n^{-0.5}).$$

Lastly, by Lemma 6(ii), condition  $(B_1)$ ,  $s_n + q_n = o(n^{1/3})$ , and (A.57), it holds that

$$\begin{aligned} \|\mathbf{R}_{n,k}\|_2 &\leq \sqrt{s_{nk} + q_n} \max_{1 \leq j \leq s_{nk} + q_n} (\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k}^\top \mathbb{W}_k^\top \text{diag}(|\mathbf{v}_{3,k}(\tilde{\boldsymbol{\theta}}_k)| \circ |\mathbb{W}_{k,j}|) \mathbb{W}_k (\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} \\ &\leq (s_n + q_n)^{3/2} N^{-1} \max_{j \in \hat{\mathbf{T}}_k} \lambda_{\max}\{\mathbf{W}_{\hat{\mathbf{T}}_k}^\top \text{diag}(|\mathbf{v}_3(\tilde{\boldsymbol{\theta}}_k)| \circ |\mathbf{W}_j|) \mathbf{W}_{\hat{\mathbf{T}}_k}\} \\ &= o_P(n^{1/2}). \end{aligned} \quad (\text{A.66})$$

This, combined with Lemma 7(i)–(ii) and  $\|\Delta_n\|_2 = O(\sqrt{m_n/n})$ , yields that

$$\|\mathbf{F}_{k,4}\|_2 \leq 0.5N^{-1}\|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)\|_2\{1 + \|\tilde{\mathbf{A}}_k^\top \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\|_2 \cdot \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*)\|_2\}\|\mathbf{R}_{n,k}\|_2 = o_P(n^{-1/2}).$$

(iii)–(iv). The asymptotic representation is obtained from (i)–(ii). To complete the proofs, it suffices to show  $\|(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{0,k})_{\hat{\mathbf{T}}_k}\|_2 = O_P(\sqrt{m_n/n})$  and results (iv). By Lemma 7(i)–(ii) and  $\|\Delta_n\|_2 = O(\sqrt{m_n/n})$ , we have  $\|\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\|_2 \leq \|\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\|_2 = O(1)$  and

$$\|\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\Delta_n\|_2 \leq \|\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \cdot \|\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\|_2 \cdot \|\Delta_n\|_2 = O_P(\sqrt{m_n/n}). \quad (\text{A.67})$$

This, combined with  $E\|\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2^2 = Nm_n\phi$ , yields that

$$\begin{aligned} & \|\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2 \\ & \leq \|\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\|_2 \cdot \|\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2 = O_P(\sqrt{nm_n}). \end{aligned} \quad (\text{A.68})$$

Note that  $\|\mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 = O(1)$ , it follows from (A.67)–(A.68) that  $\|(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{0,k})_{\hat{\mathbf{T}}_k}\|_2 = O_P(\sqrt{m_n/n})$ .  $\square$

**Lemma 9.** Let  $\mathbf{U}^{(k)}(\boldsymbol{\theta}) = N^{-1}\partial Q_n^{(k)}(\boldsymbol{\theta})/\partial\boldsymbol{\gamma}$ . Under conditions  $(A_1)$ – $(A_5)$  and  $(B_1)$ . If  $\frac{s_n+q_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), and there are two constant numbers  $C_1, C_2 \in (0, \infty)$ , independent of  $n$ , for which  $\|\mathbf{A}\mathbf{A}^\top\|_{\text{op}} \leq C_1$  and  $\|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_{\text{op}} \leq C_2$ , then under  $H_{a,2}^{(n)}$ , we have

$$(i) \quad \mathbf{U}^{(k)}(\hat{\boldsymbol{\theta}}_{0,k}) = N^{-1}\mathbf{A}^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{A}^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\Delta_n + \mathbf{r}_{n,4}^{(k)}, \text{ where } \|\mathbf{r}_{n,4}^{(k)}\|_2 = o_P(n^{-1/2}).$$

$$(ii) \quad \|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{U}^{(k)}(\hat{\boldsymbol{\theta}}_{0,k})\|_2 = O_P(\sqrt{m_n/n}).$$

**Proof of Lemma 9.** (i). Let  $\mathbf{e}_k = (\mathbf{0}_{q_n \times s_{nk}}, \mathbb{I}_{q_n})$ . Then  $\mathbf{e}_k\tilde{\mathbf{A}}_k^\top = \mathbf{A}^\top$ . By (A.61), we have

$$\begin{aligned} N^{-1}\mathbf{U}^{(k)}(\hat{\boldsymbol{\theta}}_{0,k}) &= N^{-1}\mathbf{e}_k\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + N^{-1}\mathbf{e}_k\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} + 0.5N^{-1}\mathbf{e}_k\mathbf{R}_{n,k} \\ &= N^{-1}\mathbf{e}_k\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) - \mathbf{e}_k\mathbf{I}_k(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} + 0.5N^{-1}\mathbf{e}_k\mathbf{R}_{n,k} \\ &\quad + \mathbf{e}_k\{N^{-1}\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{I}_k(\boldsymbol{\theta}^*)\}(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k}. \end{aligned}$$

Then, by Lemma 8(ii),

$$\begin{aligned} N^{-1}\mathbf{U}^{(k)}(\hat{\boldsymbol{\theta}}_{0,k}) &= N^{-1}\mathbf{e}_k\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) - \mathbf{e}_k\mathbf{I}_k(\boldsymbol{\theta}^*)\{N^{-1}\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\Delta_n \\ &\quad - N^{-1}\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{r}_{n,2}^{(k)}\} + 0.5N^{-1}\mathbf{e}_k\mathbf{R}_{n,k} \\ &\quad + \mathbf{e}_k\{N^{-1}\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{I}_k(\boldsymbol{\theta}^*)\}(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} \\ &= N^{-1}\mathbf{A}^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\tilde{\mathbf{A}}_k\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{A}^\top \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\Delta_n - \mathbf{e}_k\mathbf{I}_k(\boldsymbol{\theta}^*)\mathbf{r}_{n,2}^{(k)} \\ &\quad + \mathbf{e}_k\{N^{-1}\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{I}_k(\boldsymbol{\theta}^*)\}(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} + 0.5N^{-1}\mathbf{e}_k\mathbf{R}_{n,k}, \end{aligned}$$

where  $\|\mathbf{R}_{n,k}\|_2 = o_P(n^{1/2})$  and  $\|\mathbf{r}_{n,2}^{(k)}\|_2 = o_P(n^{-1/2})$ . Let

$$\mathbf{r}_{n,4}^{(k)} = -\mathbf{e}_k \mathbf{I}_k(\boldsymbol{\theta}^*) \mathbf{r}_{n,2}^{(k)} + 0.5N^{-1} \mathbf{e}_k \mathbf{R}_{n,k} + \mathbf{e}_k \{N^{-1} \mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{I}_k(\boldsymbol{\theta}^*)\}(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k}.$$

Then, by Lemmas 5 and 6,

$$\|\mathbf{e}_k \{N^{-1} \mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{I}_k(\boldsymbol{\theta}^*)\}(\hat{\boldsymbol{\theta}}_{0,k} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k}\|_2 = O_P((s_n + q_n)^{3/2} N^{-1}) = o_P(1/\sqrt{n}).$$

Thus,  $\|\mathbf{r}_{n,4}^{(k)}\|_2 = o_P(1/\sqrt{n})$ .

(ii). Now we prove  $\|(\mathbf{A}\mathbf{A}^\top)^{-1} \mathbf{A} \mathbf{U}^{(k)}(\hat{\boldsymbol{\theta}}_{0,k})\|_2 = O_P(\sqrt{m_n/n})$ . Since  $\|(\mathbf{A}\mathbf{A}^\top)^{-1} \mathbf{A}\|_2 = O(1)$ , we obtain that

$$\begin{aligned} & \|(\mathbf{A}\mathbf{A}^\top)^{-1} \mathbf{A} \mathbf{U}^{(k)}(\hat{\boldsymbol{\theta}}_{0,k})\|_2 \\ &= \|N^{-1} \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \boldsymbol{\Delta}_n + (\mathbf{A}\mathbf{A}^\top)^{-1} \mathbf{A} \mathbf{r}_{n,4}^{(k)}\|_2 \\ &\leq \|N^{-1} \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2 + \|\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) \boldsymbol{\Delta}_n\|_2 + o_P(1/\sqrt{n}) \\ &\leq \|\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\|_2 \cdot \|N^{-1} \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2 + \|\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\|_2 \cdot \|\boldsymbol{\Delta}_n\|_2 + o_P(1/\sqrt{n}). \end{aligned}$$

This, combined with  $E\|\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \mathbf{U}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\|_2^2 = Nm_n\phi$ , Lemma 7(i)–(ii) and  $\|\boldsymbol{\Delta}_n\|_2 = O(\sqrt{m_n/n})$ , yields that

$$\|(\mathbf{A}\mathbf{A}^\top)^{-1} \mathbf{A} \mathbf{U}^{(k)}(\hat{\boldsymbol{\theta}}_{0,k})\|_2 = O_P(\sqrt{m_n/n}). \quad \square$$

**Lemma 10.** Under conditions  $(A_1)$ – $(A_5)$  and  $(B_1)$ . Suppose  $\frac{s_n+q_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), and there are two constant numbers  $C_1, C_2 \in (0, \infty)$ , independent of  $n$ , for which  $\|\mathbf{A}\mathbf{A}^\top\|_{\text{op}} \leq C_1$  and  $\|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_{\text{op}} \leq C_2$ . If  $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 = O(\sqrt{(s_n + q_n)/n})$  and  $(\boldsymbol{\theta} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k^c} = \mathbf{0}$ , the following results hold:

- (i)  $\|\hat{\mathbf{I}}_k(\boldsymbol{\theta}) - \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 = O_P((s_n + q_n)/\sqrt{n})$  and  $\|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 = O_P((s_n + q_n)/\sqrt{n})$ ;
- (ii)  $\|\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}) - \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\|_2 = O_P((s_n + q_n)/\sqrt{n})$ ;
- (iii)  $\|\hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\|_2 = O_P((s_n + q_n)/\sqrt{n})$ ;
- (iv)  $\|\hat{\boldsymbol{\Omega}}_k^{-1/2}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_k^{-1/2}(\boldsymbol{\theta}^*)\|_2 = O_P((s_n + q_n)/\sqrt{n})$ .

**Proof of Lemma 10.** (i) By Taylor's expansion, we can get

$$\begin{aligned} \|N^{-1} \mathbb{W}_k \{\mathbf{v}_{2,k}(\boldsymbol{\theta}) - \mathbf{v}_{2,k}(\boldsymbol{\theta}^*)\} \mathbb{W}_k\|_2 &\leq \max_{\|\mathbf{a}\|_2=1} \mathbf{a}^\top \mathbb{W}_k^\top \text{diag}\{|\mathbf{v}_{3,k}(\boldsymbol{\theta}_0)| \circ |\mathbb{W}_k(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k}|\} \mathbb{W}_k \mathbf{a} / N \\ &\leq \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1 N^{-1} \max_{j \in \hat{\mathbf{T}}_k} \lambda_{\max}[\mathbf{W}_{\hat{\mathbf{T}}_k}^\top \text{diag}\{|\mathbf{v}_3(\boldsymbol{\theta}_0)| \circ |\mathbf{W}_j|\} \mathbf{W}_{\hat{\mathbf{T}}_k}] \end{aligned}$$

with  $\boldsymbol{\theta}_0 \in [\boldsymbol{\theta}, \boldsymbol{\theta}^*]$ . Since  $\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2}{\sqrt{(s_n + q_n)/n}} = O_P(1)$  and  $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_0 = O(s_n + q_n)$ , we know that  $\boldsymbol{\theta} \in \Pi_3$  and  $\frac{\sqrt{n}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_1}{s_n + q_n} = O_P(1)$ . In the same way as (A.58), we obtain that

$$\|N^{-1}\{\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}) - \mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)\}\|_2 = O_P(n^{-1/2}(s_n + q_n)). \quad (\text{A.69})$$

Using Lemma 5, together with the definition of  $\hat{\mathbf{I}}_k(\boldsymbol{\theta})$ , yields that

$$\begin{aligned} \|\hat{\mathbf{I}}_k(\boldsymbol{\theta}) - \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 &= \| -N^{-1}\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}) - \mathbf{I}_k(\boldsymbol{\theta}^*) \|_2 \\ &= \| -N^{-1}[\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}) - \mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)] - N^{-1}\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) - \mathbf{I}_k(\boldsymbol{\theta}^*) \|_2 \\ &\leq \|N^{-1}[\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}) - \mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*)]\|_2 + \|N^{-1}\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 \\ &= O_P(n^{-1/2}(s_n + q_n)). \end{aligned}$$

By condition (B<sub>1</sub>), we have

$$1/\rho_4 \leq \lambda_{\min}(\mathbf{I}_k(\boldsymbol{\theta}^*)) \leq \lambda_{\max}(\mathbf{I}_k(\boldsymbol{\theta}^*)) \leq \rho_4, \quad (\text{A.70})$$

which, combined with (A.69), leads to

$$1/(2\rho_4) \leq \lambda_{\min}(\hat{\mathbf{I}}_k(\boldsymbol{\theta})) \leq \lambda_{\max}(\hat{\mathbf{I}}_k(\boldsymbol{\theta})) \leq 2\rho_4, \quad (\text{A.71})$$

with probability going to 1, since  $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 = O_P(\sqrt{(s_n + q_n)/n})$  and  $(\boldsymbol{\theta} - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k^c} = \mathbf{0}$ . Thus,

$$\|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \leq \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta})\|_2 \|\mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 \|\hat{\mathbf{I}}_k(\boldsymbol{\theta}) - \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 = O_P(n^{-1/2}(s_n + q_n)).$$

(ii). Notice that

$$\begin{aligned} \lambda_{\min}\{\tilde{\mathbf{A}}_k \hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) \tilde{\mathbf{A}}_k^\top\} &= \inf_{\|\mathbf{a}\|_2=1} \mathbf{a}^\top \tilde{\mathbf{A}}_k \hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) \tilde{\mathbf{A}}_k^\top \mathbf{a} \\ &\geq \lambda_{\min}\{\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta})\} \inf_{\|\mathbf{a}\|_2=1} \mathbf{a}^\top \tilde{\mathbf{A}}_k \tilde{\mathbf{A}}_k^\top \mathbf{a} \\ &\geq \lambda_{\min}(\tilde{\mathbf{A}}_k \tilde{\mathbf{A}}_k^\top) \lambda_{\min}\{\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta})\} \\ &= \lambda_{\min}(\mathbf{A} \mathbf{A}^\top) / \lambda_{\max}\{\hat{\mathbf{I}}_k(\boldsymbol{\theta})\}. \end{aligned}$$

It follows that

$$\|\{\tilde{\mathbf{A}}_k \hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) \tilde{\mathbf{A}}_k^\top\}^{-1}\|_2 = \frac{1}{\lambda_{\min}\{\tilde{\mathbf{A}}_k \hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) \tilde{\mathbf{A}}_k^\top\}} \leq \lambda_{\max}\{\hat{\mathbf{I}}_k(\boldsymbol{\theta})\} / \lambda_{\min}(\mathbf{A} \mathbf{A}^\top) = O_P(1)$$

by  $\lambda_{\max}\{(\mathbf{A} \mathbf{A}^\top)^{-1}\} = O(1)$  and  $\lambda_{\max}(\hat{\mathbf{I}}_k(\hat{\boldsymbol{\theta}})) \leq 2\rho_4$ . Similarly, we get that  $\|\{\tilde{\mathbf{A}}_k(\mathbf{I}_k(\boldsymbol{\theta}))^{-1} \tilde{\mathbf{A}}_k^\top\}^{-1}\|_2 = O(1)$ . This, together with  $\lambda_{\max}(\mathbf{A} \mathbf{A}^\top) = O(1)$  and  $\|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) - \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*)\|_2 = O_P(n^{-1/2}(s_n + q_n))$

in Lemma 10(i), implies that

$$\begin{aligned}
& \|\{\tilde{\mathbf{A}}_k \hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) \tilde{\mathbf{A}}_k^\top\}^{-1} - \{\tilde{\mathbf{A}}_k \mathbf{I}_k^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k^\top\}^{-1}\|_2 \\
&= \|\{\tilde{\mathbf{A}}_k \hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k^\top\}^{-1} \tilde{\mathbf{A}}_k \{\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) - (\mathbf{I}_k(\boldsymbol{\theta}^*))^{-1}\} \tilde{\mathbf{A}}_k^\top \{\tilde{\mathbf{A}}_k (\mathbf{I}_k(\boldsymbol{\theta}))^{-1} \tilde{\mathbf{A}}_k^\top\}^{-1}\|_2 \\
&\leq \|\hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) - (\mathbf{I}_k(\boldsymbol{\theta}^*))^{-1}\|_2 \cdot \|\tilde{\mathbf{A}}_k^\top \{\tilde{\mathbf{A}}_k \hat{\mathbf{I}}_k^{-1}(\boldsymbol{\theta}) \tilde{\mathbf{A}}_k^\top\}^{-1}\|_2 \cdot \|\tilde{\mathbf{A}}_k^\top \{\tilde{\mathbf{A}}_k (\mathbf{I}_k(\boldsymbol{\theta}^*))^{-1} \tilde{\mathbf{A}}_k^\top\}^{-1}\|_2 \\
&= O_P(n^{-1/2}(s_n + q_n)).
\end{aligned}$$

(iii) Obviously,

$$\{\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*) \hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) - \hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\}^\top = -\{\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*) \hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) - \hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\}.$$

For any squared matrix  $\mathbf{C}$  with  $\mathbf{C}^\top = -\mathbf{C}$ , let  $a$  and  $\mathbf{v}$  be the eigenvalue and corresponding eigenvector of  $\mathbf{C}$ . Then

$$a^2 \|\mathbf{v}\|_2^2 = \mathbf{v}^\top \mathbf{C}^\top \mathbf{C} \mathbf{v} = -\mathbf{v}^\top \mathbf{C} \mathbf{C} \mathbf{v} = -a^2 \|\mathbf{v}\|_2^2. \quad (\text{A.72})$$

Thus, all eigenvalues of  $\mathbf{C}^\top \mathbf{C}$  are 0. Hence,

$$\|\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*) \hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) - \hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\|_2 = 0. \quad (\text{A.73})$$

Note that the identity

$$\begin{aligned}
(\hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) + \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*))(\hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)) &= \hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}) - \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*) + \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*) \hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) \\
&\quad - \hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*).
\end{aligned} \quad (\text{A.74})$$

For any positive definite matrix  $\mathbf{C}_1$  and square matrix  $\mathbf{C}_2$ , we have

$$\|\mathbf{C}_1 \mathbf{C}_2\|_2^2 = \sup_{\|\mathbf{a}\|_2=1} \mathbf{a}^\top \mathbf{C}_2^\top \mathbf{C}_1 \mathbf{C}_2 \mathbf{a} \geq \lambda_{\min}(\mathbf{C}_1^2) \|\mathbf{C}_2\|_2^2.$$

Hence,  $\|\mathbf{C}_1 \mathbf{C}_2\|_2 \geq \lambda_{\min}(\mathbf{C}_1) \|\mathbf{C}_2\|_2$ . Then, by (A.73)-(A.74),

$$\begin{aligned}
\lambda_{\min}(\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)) \|\hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\|_2 &\leq \|(\hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) + \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*))(\hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*))\|_2 \\
&\leq \|\hat{\boldsymbol{\Omega}}_k(\boldsymbol{\theta}) - \boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\|_2.
\end{aligned} \quad (\text{A.75})$$

By the definition of  $\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)$ ,  $\lambda_{\max}(\mathbf{A} \mathbf{A}^\top) = O(1)$ , and (A.70), It is straightforward to check that

$$\lambda_{\max}(\boldsymbol{\Omega}_k^{-1}(\boldsymbol{\theta}^*)) \leq \lambda_{\max}(\mathbf{A} \mathbf{A}^\top) / \lambda_{\min}(\mathbf{I}_k(\boldsymbol{\theta}^*)) = O(1).$$

Then, by result (ii) and (A.75), we have  $\|\hat{\boldsymbol{\Omega}}_k^{1/2}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\|_2 = O_P((s_n + q_n)n^{-1/2})$ .

(iv) Similar to (A.75), we can show that

$$\begin{aligned}\lambda_{\min}(\hat{\Omega}_k^{-1/2}(\theta^*))\|\hat{\Omega}_k^{-1/2}(\theta) - \Omega_k^{-1/2}(\theta^*)\|_2 &\leq \|(\hat{\Omega}_k^{-1/2}(\theta) + \Omega_k^{-1/2}(\theta^*))(\hat{\Omega}_k^{-1/2}(\theta) - \Omega_k^{-1/2}(\theta^*))\|_2 \\ &\leq \|\hat{\Omega}_k^{-1}(\theta) - \Omega_k^{-1}(\theta^*)\|_2.\end{aligned}$$

That is,

$$\begin{aligned}\|\hat{\Omega}_k^{-1/2}(\theta) - \Omega_k^{-1/2}(\theta^*)\|_2 &\leq \|\tilde{\mathbf{A}}_k\{\mathbf{I}_k^{-1}(\theta^*) - \hat{\mathbf{I}}_k^{-1}(\theta)\}\tilde{\mathbf{A}}_k^\top\|_2/\sqrt{\lambda_{\min}(\tilde{\mathbf{A}}_k\mathbf{I}_k^{-1}(\theta^*)\tilde{\mathbf{A}}_k)} \\ &\leq \lambda_{\max}(\mathbf{I}_k(\theta^*))\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)\|\mathbf{I}_k^{-1}(\theta^*) - \hat{\mathbf{I}}_k^{-1}(\theta)\|_2/\lambda_{\min}(\mathbf{A}\mathbf{A}^\top).\end{aligned}$$

Hence, by result (i),  $\lambda_{\max}(\mathbf{A}\mathbf{A}^\top) = O(1)$ ,  $\lambda_{\min}((\mathbf{A}\mathbf{A}^\top)^{-1}) = O(1)$ , and  $\lambda_{\max}(\mathbf{I}_k(\theta^*)) \leq \rho_4$  in (A.70), we get

$$\|\hat{\Omega}_k^{-1/2}(\theta) - \Omega_k^{-1/2}(\theta^*)\|_2 = O_P((s_n + q_n)n^{-1/2}). \quad \square$$

**Lemma 11.** Let  $\omega_{n,k} = N^{-1/2}\tilde{\mathbf{A}}_k\mathbf{I}_k^{-1}(\theta^*)U_{\hat{\mathbf{T}}_k}^{(k)}(\theta^*) + N^{1/2}\Delta_n$ . Under conditions (A<sub>1</sub>)-(A<sub>5</sub>) and (B<sub>1</sub>). If  $\frac{s_n+q_n}{n^{1/3}} \rightarrow 0$  ( $n \rightarrow \infty$ ), and there are two constant numbers  $C_1, C_2 \in (0, \infty)$ , independent of  $n$ , for which  $\|\mathbf{A}\mathbf{A}^\top\|_{\text{op}} \leq C_1$  and  $\|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_{\text{op}} \leq C_2$ , then the following results hold:

- (i)  $T = (2\phi)^{-1}\|\sum_{k=1}^2\Omega_k^{1/2}(\theta^*)\omega_{n,k}\|_2^2 + o_P(m_n)$ , where  $T$  is  $T_{n,\mathbf{L}}$ ,  $T_{n,\mathbf{W}}$  or  $T_{n,\mathbf{S}}$ ;
- (ii)  $2\{Q_n^{(k)}(\hat{\theta}_k) - Q_n^{(k)}(\hat{\theta}_{0,k})\} = \|\Omega_k^{1/2}(\theta^*)\omega_{n,k}\|_2^2 + o_P(m_n)$ ;
- (iii)  $\sqrt{N}\zeta_{n,k} = \Omega_k^{1/2}(\theta^*)\omega_{n,k} + o_P(1)$ ;
- (iv) if condition (A<sub>6</sub>) holds, then

$$T_{n,2} = \|\{\tilde{\mathbf{A}}_*\mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\theta^*)\tilde{\mathbf{A}}_*^\top\}^{-1}[n^{-1/2}\tilde{\mathbf{A}}_*\mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\theta^*)\mathbf{W}_{\hat{\mathbf{T}}}^\top\mathbf{v}_1(\theta^*) + n^{1/2}\Delta_n]\|_2^2 + o_P(m_n),$$

where  $\tilde{\mathbf{A}}_* = (\mathbf{0}_{m_n \times \|\hat{\mathbf{s}}\|_0}, \mathbf{A})$ .

**Proof of Lemma 11.** (i) We divide the proof into three steps for (1)  $T = T_{n,\mathbf{L}}$ , (2)  $T = T_{n,\mathbf{W}}$ , and (3)  $T = T_{n,\mathbf{S}}$ .

Step (1). Let  $\mathbf{L}_k = (\mathbf{A}\mathbf{A}^\top)^{-1}\tilde{\mathbf{A}}_k$ ,  $\hat{\mathbf{C}}_k = \hat{\Omega}_k^{-1/2}(\hat{\theta}_k)\mathbf{L}_k\hat{\mathbf{I}}_k(\hat{\theta}_k)(\hat{\theta}_k - \hat{\theta}_{0,k})_{\hat{\mathbf{T}}_k}$ , and  $\mathbf{C}_k = \Omega_k^{-1/2}(\theta^*)\mathbf{L}_k\mathbf{I}_k(\theta^*)(\hat{\theta}_k - \hat{\theta}_{0,k})_{\hat{\mathbf{T}}_k}$ . It follows from  $\mathbf{A}\hat{\mathbf{I}}_k^{22}(\hat{\theta}_k)\mathbf{A}^\top = \tilde{\mathbf{A}}_k\hat{\mathbf{I}}_k(\hat{\theta}_k)\tilde{\mathbf{A}}_k^\top$  and the definition of  $T_{n,\mathbf{L}}$  that

$$T_{n,\mathbf{L}} = \frac{N}{2\hat{\phi}_R} \left\| \sum_{k=1}^2 \hat{\Omega}_k^{-1/2}(\hat{\theta}_k)\mathbf{L}_k\hat{\mathbf{I}}_k(\hat{\theta}_k)(\hat{\theta}_k - \hat{\theta}_{0,k})_{\hat{\mathbf{T}}_k} \right\|_2^2 = \frac{N}{2\hat{\phi}_R} \|\mathbf{C}_1 + \mathbf{C}_2\|_2^2 + L_{n,1}, \quad (\text{A.76})$$

where  $L_{n,1} = \frac{N}{2\hat{\phi}_R} (\|\hat{\mathbf{C}}_1 + \hat{\mathbf{C}}_2\|_2 + \|\mathbf{C}_1 + \mathbf{C}_2\|_2) (\|\hat{\mathbf{C}}_1 + \hat{\mathbf{C}}_2\|_2 - \|\mathbf{C}_1 + \mathbf{C}_2\|_2)$ . Now we prove that  $L_{n,1} = o_P(m_n)$ . By  $\hat{\mathbf{\Omega}}_k(\hat{\boldsymbol{\theta}}_k) = \{\tilde{\mathbf{A}}_k \hat{\mathbf{I}}_k(\hat{\boldsymbol{\theta}}_k) \tilde{\mathbf{A}}_k^\top\}^{-1}$ ,  $\mathbf{\Omega}_k(\boldsymbol{\theta}^*) = \{\tilde{\mathbf{A}}_k \mathbf{I}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k^\top\}^{-1}$ ,  $\|\tilde{\mathbf{A}}_k^\top \mathbf{L}_k\|_2 = \|\mathbf{A}(\mathbf{A}^\top \mathbf{A}^\top)^{-1} \mathbf{A}\|_2 = 1$ , (A.70)-(A.71), Lemma 8(iii), and Lemma 10(i)&(iv), we obtain that

$$\|\mathbf{\Omega}_k^{-1/2}(\boldsymbol{\theta}^*) \mathbf{L}_k \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2^2 \leq \|\mathbf{L}_k^\top \tilde{\mathbf{A}}_k \mathbf{I}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k^\top \mathbf{L}_k\|_2 \cdot \|\mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 = O(1), \quad (\text{A.77})$$

$$\begin{aligned} \|\hat{\mathbf{C}}_k\|_2 + \|\mathbf{C}_k\|_2 &\leq \{\|\hat{\mathbf{\Omega}}_k^{-1/2}(\hat{\boldsymbol{\theta}}_k) \mathbf{L}_k\|_2 \cdot \|\hat{\mathbf{I}}_k(\hat{\boldsymbol{\theta}}_k)\|_2 + \|\mathbf{\Omega}_k^{-1/2}(\boldsymbol{\theta}^*) \mathbf{L}_k\|_2 \cdot \|\mathbf{I}_k(\hat{\boldsymbol{\theta}}_k)\|_2\} \cdot \|(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{0,k})_{\hat{\mathbf{T}}_k}\|_2 \\ &= \{\|\mathbf{L}_k^\top \tilde{\mathbf{A}}_k \hat{\mathbf{I}}_k(\hat{\boldsymbol{\theta}}_k) \tilde{\mathbf{A}}_k^\top \mathbf{L}_k\|_2^{1/2} + \|\mathbf{L}_k^\top \tilde{\mathbf{A}}_k \mathbf{I}_k(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k^\top \mathbf{L}_k\|_2^{1/2}\} O_P(\sqrt{m_n/n}) \\ &= O_P(\sqrt{m_n/n}), \end{aligned}$$

and

$$\begin{aligned} \|\hat{\mathbf{C}}_k - \mathbf{C}_k\|_2 &\leq \|\hat{\mathbf{\Omega}}_k^{-1/2}(\hat{\boldsymbol{\theta}}_k) - \mathbf{\Omega}_k^{-1/2}(\boldsymbol{\theta}^*)\|_2 \cdot \|\mathbf{L}_k\|_2 \cdot \|\hat{\mathbf{I}}_k(\hat{\boldsymbol{\theta}}_k)\|_2 \cdot \|(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{0,k})_{\hat{\mathbf{T}}_k}\|_2 \\ &\quad + \|\mathbf{\Omega}_k^{-1/2}(\boldsymbol{\theta}^*) \mathbf{L}_k\|_2 \cdot \|\hat{\mathbf{I}}_k(\hat{\boldsymbol{\theta}}_k) - \mathbf{I}_k(\boldsymbol{\theta}^*)\|_2 \cdot \|(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{0,k})_{\hat{\mathbf{T}}_k}\|_2 \\ &= O_P((s_n + q_n)n^{-1/2}) O_P(\sqrt{m_n/n}), \end{aligned} \quad (\text{A.78})$$

for  $k = 1, 2$ . Applying the triangle inequality and  $s_n + q_n = o(n^{1/3})$ , we establish that

$$\begin{aligned} |L_{n,1}| &\leq \frac{N}{2\hat{\phi}_R} (\|\hat{\mathbf{C}}_1 + \hat{\mathbf{C}}_2\|_2 + \|\mathbf{C}_1 + \mathbf{C}_2\|_2) \left\| \sum_{j=1}^2 (\hat{\mathbf{C}}_j - \mathbf{C}_j) \right\|_2 \\ &\leq \frac{N}{2\hat{\phi}_R} \sum_{j=1}^2 (\|\hat{\mathbf{C}}_1 + \hat{\mathbf{C}}_2\|_2 + \|\mathbf{C}_1 + \mathbf{C}_2\|_2) \|\hat{\mathbf{C}}_j - \mathbf{C}_j\|_2 \\ &\leq \frac{N}{2\hat{\phi}_R} \sum_{k,j=1}^2 (\|\hat{\mathbf{C}}_k\|_2 + \|\mathbf{C}_k\|_2) \|\hat{\mathbf{C}}_j - \mathbf{C}_j\|_2 \\ &= o_P(m_n). \end{aligned} \quad (\text{A.79})$$

This, combined with (A.76), yields that

$$T_{n,\mathbf{L}} = \frac{N}{2\hat{\phi}_R} \left\| \sum_{k=1}^2 \mathbf{\Omega}_k^{-1/2}(\boldsymbol{\theta}^*) \mathbf{L}_k \mathbf{I}_k(\boldsymbol{\theta}^*) (\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{0,k})_{\hat{\mathbf{T}}_k} \right\|_2^2 + o_P(m_n). \quad (\text{A.80})$$

Applying Lemma 8(iii), we obtain that

$$T_{n,\mathbf{L}} = \frac{1}{2\hat{\phi}_R} \left\| \sum_{k=1}^2 \mathbf{\Omega}_k^{1/2}(\boldsymbol{\theta}^*) \boldsymbol{\omega}_{n,k} + N^{1/2} \mathbf{\Omega}_k^{-1/2}(\boldsymbol{\theta}^*) \mathbf{L}_k \mathbf{I}_k(\boldsymbol{\theta}^*) \mathbf{r}_{n,3}^{(k)} \right\|_2^2 + o_P(m_n),$$

where  $\|\mathbf{r}_{n,3}^{(k)}\|_2 = o_P(n^{-1/2})$ . Using Lemma 8(iv) and (A.77), we get

$$\begin{aligned} \|\mathbf{\Omega}_k^{1/2}(\boldsymbol{\theta}^*) \boldsymbol{\omega}_{n,k}\|_2 &= \|\{\mathbf{I}_k(\boldsymbol{\theta}^*)\}^{-1/2} \tilde{\mathbf{A}}_k^\top \mathbf{\Omega}_k(\boldsymbol{\theta}^*) \boldsymbol{\omega}_{n,k}\|_2 = O_P(\sqrt{m_n}); \\ \|N^{1/2} \mathbf{\Omega}_k^{-1/2}(\boldsymbol{\theta}^*) \mathbf{L}_k \mathbf{I}_k(\boldsymbol{\theta}^*) \mathbf{r}_{n,3}^{(k)}\|_2 &= o_P(1). \end{aligned} \quad (\text{A.81})$$

Similar to (A.80), using (A.81), we arrive at

$$\begin{aligned} T_{n,\mathbf{L}} &= \frac{1}{2} \left\| \sum_{k=1}^2 \Omega_k^{1/2}(\boldsymbol{\theta}^*) \boldsymbol{\omega}_{n,k} \right\|_2^2 + o_P(m_n) \\ &= \frac{1}{2} \left\| \sum_{k=1}^2 \Omega_k^{1/2}(\boldsymbol{\theta}^*) \boldsymbol{\omega}_{n,k} \right\|_2^2 + o_P(m_n). \end{aligned} \quad (\text{A.82})$$

Step (2). Under  $H_{a,2}^{(n)}$ , we have  $\mathbf{A}(\boldsymbol{\gamma}^* - \mathbf{r}) = \boldsymbol{\Delta}_n$ . Then

$$\mathbf{A}(\hat{\boldsymbol{\gamma}}_k - \mathbf{r}) = \mathbf{A}(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}^* + \boldsymbol{\gamma}^* - \mathbf{r}) = \tilde{\mathbf{A}}_k(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k} + \boldsymbol{\Delta}_n.$$

Note that  $\|\boldsymbol{\Delta}_n\|_2 = O(\sqrt{m_n/n})$ ,  $\|\tilde{\mathbf{A}}_k\|_2 = \|\mathbf{A}\|_2 = O(1)$ , and  $\|(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)_{\hat{\mathbf{T}}_k}\|_2 = O_P(\sqrt{(s_n + q_n)/n})$  in Lemma 6(i). It follows that

$$\sqrt{\frac{n}{s_n + q_n}} \|\mathbf{A}(\hat{\boldsymbol{\gamma}}_k - \mathbf{r})\|_2 = O_P(1). \quad (\text{A.83})$$

This, combined Lemma 7(ii) and Lemma 10(iii), implies that

$$\sqrt{\frac{n}{s_n + q_n}} \left\| \hat{\Omega}_k^{1/2}(\hat{\boldsymbol{\theta}}_k) \mathbf{A}(\hat{\boldsymbol{\gamma}}_k - \mathbf{r}) \right\|_2 = O_P(1); \quad \sqrt{\frac{n}{s_n + q_n}} \left\| \Omega_k^{1/2}(\boldsymbol{\theta}^*) \mathbf{A}(\hat{\boldsymbol{\gamma}}_k - \mathbf{r}) \right\|_2 = O_P(1). \quad (\text{A.84})$$

It follows from Lemma 8(i) that

$$\begin{aligned} \Omega_k^{1/2}(\boldsymbol{\theta}^*) \mathbf{A}(\hat{\boldsymbol{\gamma}}_k - \mathbf{t}) &= \Omega_k^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \{N \mathbf{I}_k(\boldsymbol{\theta}^*)\}^{-1} U_{\hat{\mathbf{T}}_k}^{(k)}(\boldsymbol{\theta}^*) + \Omega_k^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \mathbf{r}_{n,1}^{(k)} + \Omega_k^{1/2}(\boldsymbol{\theta}^*) \boldsymbol{\Delta}_n \\ &= N^{-1/2} \Omega_k^{1/2}(\boldsymbol{\theta}^*) \boldsymbol{\omega}_{n,k} + \Omega_k^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \mathbf{r}_{n,1}^{(k)}. \end{aligned} \quad (\text{A.85})$$

Similar to (A.80), by Lemma 10(iii) and (A.83)-(A.84), we can show that

$$\begin{aligned} T_{n,\mathbf{W}} &= \frac{N}{2} \left\| \sum_{k=1}^2 \hat{\Omega}_k^{1/2}(\hat{\boldsymbol{\theta}}_k) \mathbf{A}(\hat{\boldsymbol{\gamma}}_k - \mathbf{r}) \right\|_2^2 \\ &= \frac{N}{2} \left\| \sum_{k=1}^2 \Omega_k^{1/2}(\boldsymbol{\theta}^*) \mathbf{A}(\hat{\boldsymbol{\gamma}}_k - \mathbf{r}) \right\|_2^2 + o_P(m_n). \end{aligned}$$

Thus, by (A.85), we have

$$T_{n,\mathbf{W}} = \frac{1}{2} \left\| \sum_{k=1}^2 \Omega_k^{1/2}(\boldsymbol{\theta}^*) \boldsymbol{\omega}_{n,k} + N^{1/2} \Omega_k^{1/2}(\boldsymbol{\theta}^*) \tilde{\mathbf{A}}_k \mathbf{r}_{n,1}^{(k)} \right\|_2^2 + o_P(m_n).$$

Then, similar to Step (1), we can show that

$$T_{n,\mathbf{W}} = \phi \left\| \sum_{k=1}^2 \Omega_k^{1/2}(\boldsymbol{\theta}^*) \boldsymbol{\omega}_{n,k} \right\|_2^2 + o_P(m_n). \quad (\text{A.86})$$

Step (3). Using  $\lambda_{\max}\{(\mathbf{A}\mathbf{A}^\top)^{-1}\} = O(1)$ ,  $\lambda_{\max}(\mathbf{A}\mathbf{A}^\top) = O(1)$ ,  $\|\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\|_2 = O(1)$ , (A.81), Lemma 7(ii), and Lemma 9, we obtain that

$$(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{U}(\hat{\boldsymbol{\theta}}_{0,k}) = N^{-1/2}\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\boldsymbol{\omega}_{n,k} + (\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{r}_{n,4}^{(k)},$$

and

$$\begin{aligned} \|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{U}(\hat{\boldsymbol{\theta}}_{0,k})\|_2 &\leq \|N^{-1/2}\boldsymbol{\Omega}_k(\boldsymbol{\theta}^*)\boldsymbol{\omega}_{n,k}\|_2 + \|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{r}_{n,4}^{(k)}\|_2 \\ &\leq \|N^{-1/2}\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\|_2 \cdot \|\boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\boldsymbol{\omega}_{n,k}\|_2 + \|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_2 \cdot \|\mathbf{r}_{n,4}^{(k)}\|_2 \\ &= O_P(\sqrt{m_n/n}), \end{aligned}$$

where the second line is obtained by Hölder inequality and

$$\|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\|_2^2 = \|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}\|_2 = \|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_2.$$

This, combined with (A.70)-(A.71),  $\hat{\boldsymbol{\Omega}}_k(\hat{\boldsymbol{\theta}}_{0,k}) = \{\tilde{\mathbf{A}}_k\hat{\mathbf{I}}_k(\hat{\boldsymbol{\theta}}_{0,k})\tilde{\mathbf{A}}_k^\top\}^{-1}$ , and  $\lambda_{\max}\{(\mathbf{A}\mathbf{A}^\top)^{-1}\} = O(1)$ , yields that

$$\begin{aligned} \|\hat{\boldsymbol{\Omega}}_k^{-1/2}(\hat{\boldsymbol{\theta}}_{0,k})(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{U}(\hat{\boldsymbol{\theta}}_{0,k})\|_2 &\leq \|\hat{\boldsymbol{\Omega}}_k^{-1}(\hat{\boldsymbol{\theta}}_{0,k})\|_2^{1/2} \|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{U}(\hat{\boldsymbol{\theta}}_{0,k})\|_2 \\ &= O_P(\sqrt{m_n/n}) \|\hat{\mathbf{I}}_k^{-1}(\hat{\boldsymbol{\theta}}_{0,k})\|_2^{1/2} \cdot \|(\mathbf{A}\mathbf{A}^\top)^{-1}\|_2^{1/2} \\ &= O_P(\sqrt{m_n/n}). \end{aligned}$$

Similarly,  $\|\boldsymbol{\Omega}_k^{-1/2}(\boldsymbol{\theta}^*)(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{U}(\hat{\boldsymbol{\theta}}_{0,k})\|_2 = O_P(\sqrt{m_n/n})$ . Similar to (A.80), it holds that

$$\begin{aligned} T_{n,s} &= \frac{N}{2} \left\| \sum_{k=1}^2 \hat{\boldsymbol{\Omega}}_k^{-1/2}(\hat{\boldsymbol{\theta}}_{0,k})(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{U}(\hat{\boldsymbol{\theta}}_{0,k}) \right\|_2^2 \\ &= \frac{N}{2} \left\| \sum_{k=1}^2 \boldsymbol{\Omega}_k^{-1/2}(\boldsymbol{\theta}^*)(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\mathbf{U}(\hat{\boldsymbol{\theta}}_{0,k}) \right\|_2^2 + o_P(m_n) \\ &= \frac{1}{2} \left\| \sum_{k=1}^2 \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\boldsymbol{\omega}_{n,k} + N^{1/2}\boldsymbol{\Omega}_k^{-1/2}(\boldsymbol{\theta}^*)(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}^\top\mathbf{r}_{n,4}^{(k)} \right\|_2^2 + o_P(m_n). \end{aligned}$$

With the same argument as used in Step (1), we can also show that

$$T_{n,s} = (2\phi)^{-1} \left\| \sum_{k=1}^2 \boldsymbol{\Omega}_k^{1/2}(\boldsymbol{\theta}^*)\boldsymbol{\omega}_{n,k} \right\|_2^2 + o_P(m_n).$$

(ii) For the  $k$ -th dataset, combining the mean value theorem with Lemma 8, then

$$\begin{aligned} Q_n^{(k)}(\hat{\boldsymbol{\theta}}_{0,k}) - Q_n^{(k)}(\hat{\boldsymbol{\theta}}_k) &= \{U_{n,\hat{\mathbf{T}}_k}^{(k)}(\hat{\boldsymbol{\theta}}_k)\}^\top (\hat{\boldsymbol{\theta}}_{0,k} - \hat{\boldsymbol{\theta}}_k)_{\hat{\mathbf{T}}_k} \\ &\quad + 0.5 \{(\hat{\boldsymbol{\theta}}_{0,k} - \hat{\boldsymbol{\theta}}_k)_{\hat{\mathbf{T}}_k}\}^\top \mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\tilde{\boldsymbol{\theta}}_k)(\hat{\boldsymbol{\theta}}_{0,k} - \hat{\boldsymbol{\theta}}_k)_{\hat{\mathbf{T}}_k}, \end{aligned} \quad (\text{A.87})$$

with  $\tilde{\theta}_k \in [\hat{\theta}_{0,k}, \hat{\theta}_k]$ . Since  $\hat{\theta}_k = \arg \max\{Q_n^{(k)}(\theta), \text{ s.t. } \beta_{\hat{\mathbf{S}}_k} = \mathbf{0}\}$ , we get  $U_{n, \hat{\mathbf{T}}_k}^{(k)}(\hat{\theta}_k) = \mathbf{0}$ . By Lemma 6(i), we have  $\|\tilde{\theta}_k - \theta^*\|_2 \leq \|\hat{\theta}_k - \theta^*\|_2 = O_P(\sqrt{(s_n + q_n)/n})$ . Recall that  $(\tilde{\theta}_k - \theta^*)_{\hat{\mathbf{T}}_k^c} = (\hat{\theta}_k - \theta^*)_{\hat{\mathbf{T}}_k^c} = \mathbf{0}$  in Lemma 8(i). Then, by Lemma 8(iii) and Lemma 10(i), we get

$$N\{(\hat{\theta}_{0,k} - \hat{\theta}_k)_{\hat{\mathbf{T}}_k}\}^\top \{\mathbf{H}_{\hat{\mathbf{T}}_k}^{(k)}(\tilde{\theta}_k)/N + \mathbf{I}_k(\theta^*)\}(\hat{\theta}_{0,k} - \hat{\theta}_k)_{\hat{\mathbf{T}}_k} = O_P\{m_n(s_n + q_n)n^{-1/2}\}.$$

Note that  $\|\{\mathbf{I}_k(\theta^*)\}^{-1/2} \tilde{\mathbf{A}}_k^\top \Omega_k(\theta^*) \omega_{n,k}\|_2^2 = \|\Omega_k^{1/2}(\theta^*) \omega_{n,k}\|_2^2$ . It follows from Lemma 8(iii) and (A.87) that

$$\begin{aligned} & 2Q_n^{(k)}(\hat{\theta}_k) - 2Q_n^{(k)}(\hat{\theta}_{0,k}) \\ &= N\{(\hat{\theta}_{0,k} - \hat{\theta}_k)_{\hat{\mathbf{T}}_k}\}^\top \mathbf{I}_k(\theta^*)(\hat{\theta}_{0,k} - \hat{\theta}_k)_{\hat{\mathbf{T}}_k} + o_P(m_n) \\ &= \|\{\mathbf{I}_k(\theta^*)\}^{-1/2} \tilde{\mathbf{A}}_k^\top \Omega_k(\theta^*) \omega_{n,k} + N^{1/2}\{\mathbf{I}_k(\theta^*)\}^{1/2} \mathbf{r}_{n,3}^{(k)}\|_2^2 + o_P(m_n) \\ &= \|\Omega_k^{1/2}(\theta^*) \omega_{n,k}\|_2^2 + N\|\{\mathbf{I}_k(\theta^*)\}^{1/2} \mathbf{r}_{n,3}^{(k)}\|_2^2 + 2N^{1/2}\{\mathbf{r}_{n,3}^{(k)}\}^\top \tilde{\mathbf{A}}_k^\top \Omega_k(\theta^*) \omega_{n,k} + o_P(m_n) \\ &= \|\Omega_k^{1/2}(\theta^*) \omega_{n,k}\|_2^2 + o_P(m_n), \end{aligned}$$

where the last equality follows from  $\|\{\mathbf{I}_k(\theta^*)\}^{1/2} \mathbf{r}_{n,3}^{(k)}\|_2 = o_P(n^{-1/2})$  and  $\|\{\mathbf{I}_k(\theta^*)\}^{-1/2} \tilde{\mathbf{A}}_k^\top \Omega_k(\theta^*) \omega_{n,k}\|_2 = O_P(\sqrt{m_n})$  in Lemma 8(iv).

(iii) By the definition of  $\zeta_{n,k}$  in Theorem 4, (A.78), (A.81), Lemma 8(iii) and  $m_n \leq s_n + q_n = o(n^{1/3})$ , we obtain that

$$\begin{aligned} \sqrt{N}\zeta_{n,k} &= \sqrt{N}\Omega_k^{-1/2}(\theta^*) \mathbf{L}_k \mathbf{I}_k(\theta^*)(\hat{\theta}_k - \hat{\theta}_{0,k})_{\hat{\mathbf{T}}_k} + \sqrt{N}(\hat{\mathbf{C}}_k - \mathbf{C}_k) \\ &= \Omega_k^{1/2}(\theta^*) \omega_{n,k} + N^{1/2}\Omega_k^{-1/2}(\theta^*) \mathbf{L}_k \mathbf{I}_k(\theta^*) \mathbf{r}_{n,3}^{(k)} + o_P(1) \\ &= \Omega_k^{1/2}(\theta^*) \omega_{n,k} + o_P(1). \end{aligned}$$

(iv) By Theorem 1, we know that  $P(\hat{\mathbf{S}} = \mathbf{S}) \rightarrow 1$ . Similar to result (ii), it can be shown that

$$T_{n,2} = \|\{\tilde{\mathbf{A}}_* \mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\theta^*) \tilde{\mathbf{A}}_*^\top\}^{-1/2} [n^{-1/2} \tilde{\mathbf{A}}_* \mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\theta^*) \mathbf{U}_{\hat{\mathbf{T}}}(\theta^*) + n^{1/2} \Delta_n]\|_2^2 + o_P(m_n).$$

Then

$$T_{n,2} = \|\{\tilde{\mathbf{A}}_* \mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\theta^*) \tilde{\mathbf{A}}_*^\top\}^{-1/2} [n^{-1/2} \tilde{\mathbf{A}}_* \mathbf{I}_{\hat{\mathbf{T}}}^{-1}(\theta^*) \mathbf{U}_{\hat{\mathbf{T}}}(\theta^*) + n^{1/2} \Delta_n]\|_2^2 + o_P(m_n). \quad \square$$

**Lemma 12.** (Bentkus, 2004, Th1) Let  $(\mathcal{X}_i)_{i=1}^n$  be independent random vectors in  $\mathbb{R}^m$  with  $E(\mathcal{X}_i) = \mathbf{0}$ ,  $\sum_{i=1}^n \text{Cov}(\mathcal{X}_i) = \mathbf{C}^2$  and  $\mathcal{Z}$  be a Gaussian random vector for which  $E(\mathcal{Z}) = \mathbf{0}$ ,  $\text{Cov}(\mathcal{Z}) = \mathbf{C}^2$ . Choose  $c_* > 0$  for which

$$\sup_{\psi} |P(\sum_{i=1}^n \mathcal{X}_i \in \psi) - P(\mathcal{Z} \in \psi)| \leq c_*(m)^{1/4} \sum_{i=1}^n E\|\mathbf{C}^{-1} \mathcal{X}_i\|_2^3,$$

with the supremum over  $C \subset \mathbb{R}^m$  convex.

**Lemma 13.** Let  $\eta_{L,k}^2 = n\Delta_n^\top \{\mathbf{A}\mathbf{I}_k^{22}(\boldsymbol{\theta}^*)\mathbf{A}^\top\}^{-1}\Delta_n$  for  $k = 1, 2$ . Assume conditions  $(A_1)$ – $(A_5)$  and  $(B_2)$  hold. Then

- (i)  $\eta_n^2 \leq \eta_{n,o}^2$  and  $\eta_n^2 \leq \eta_{L,k}^2 \leq \eta_{n,o}^2$ ;
- (ii)  $n\| [2\{\mathbf{A}\mathbf{I}_o^{22}(\boldsymbol{\theta}^*)\mathbf{A}^\top\}^{-1/2} - \sum_{k=1}^2 \{\mathbf{A}\mathbf{I}_k^{22}(\boldsymbol{\theta}^*)\mathbf{A}^\top\}^{-1/2}] \Delta_n \|_2^2 \leq 16c^2\eta_{n,L}^2$ , where  $c = \frac{\rho^*\sqrt{1+\rho^*}}{1-\rho^*}$  and  $\rho^* \in [0, 1)$  is the constant specified in  $(B_3)$ .

**Proof of Lemma 13.** For convenience, let  $\boldsymbol{\Sigma} = \mathbf{I}(\boldsymbol{\theta}^*)$ . By the definitions of  $\eta_n^2$ ,  $\eta_{L,k}^2$ , and  $\eta_{n,o}^2$ , we obtain that

$$\begin{aligned}\eta_n^2 &= n\Delta_n^\top \{\mathbf{A}(\boldsymbol{\Sigma}_D - \boldsymbol{\Sigma}_{D,D^c}\boldsymbol{\Sigma}_{D^c}^{-1}\boldsymbol{\Sigma}_{D^c,D})^{-1}\mathbf{A}^\top\}^{-1}\Delta_n; \\ \eta_{L,k}^2 &= n\Delta_n^\top \{\mathbf{A}(\boldsymbol{\Sigma}_D - \boldsymbol{\Sigma}_{D,\hat{\mathbf{s}}_k}\boldsymbol{\Sigma}_{\hat{\mathbf{s}}_k}^{-1}\boldsymbol{\Sigma}_{\hat{\mathbf{s}}_k,D})^{-1}\mathbf{A}^\top\}^{-1}\Delta_n; \\ \eta_{n,o}^2 &= n\Delta_n^\top \{\mathbf{A}(\boldsymbol{\Sigma}_D - \boldsymbol{\Sigma}_{D,S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S,D})^{-1}\mathbf{A}^\top\}^{-1}\Delta_n.\end{aligned}$$

For any  $q_n \times q_n$  positive definite matrix  $\mathbf{E}_1$  and positive semi-definite matrix  $\mathbf{E}_2$ , we have

$$\begin{aligned}& \|(\mathbf{A}\mathbf{E}_1^{-1}\mathbf{A}^\top)^{-1/2}\mathbf{A}(\mathbf{E}_1 + \mathbf{E}_2)^{-1}\mathbf{A}^\top(\mathbf{A}\mathbf{E}_1^{-1}\mathbf{A}^\top)^{-1/2}\|_2 \\ & \leq \|(\mathbb{I}_{q_n} + \mathbf{E}_1^{-1/2}\mathbf{E}_2\mathbf{E}_1^{-1/2})^{-1}\|_2 \|\mathbf{E}_1^{-1/2}\mathbf{A}^\top(\mathbf{A}\mathbf{E}_1^{-1}\mathbf{A}^\top)^{-1/2}\|_2^2 \\ & \leq \{1 + \lambda_{\min}(\mathbf{E}_1^{-1/2}\mathbf{E}_2\mathbf{E}_1^{-1/2})\}^{-1},\end{aligned}$$

which implies that

$$\begin{aligned}& \Delta_n^\top \{\mathbf{A}(\mathbf{E}_1 + \mathbf{E}_2)^{-1}\mathbf{A}^\top\}^{-1}\Delta_n \\ & = \|[(\mathbf{A}\mathbf{E}_1^{-1}\mathbf{A}^\top)^{-1/2}\mathbf{A}(\mathbf{E}_1 + \mathbf{E}_2)^{-1}\mathbf{A}^\top(\mathbf{A}\mathbf{E}_1^{-1}\mathbf{A}^\top)^{-1/2}]^{-1/2}(\mathbf{A}\mathbf{E}_1^{-1}\mathbf{A}^\top)^{-1/2}\Delta_n\|_2^2 \\ & \geq \{1 + \lambda_{\min}(\mathbf{E}_1^{-1/2}\mathbf{E}_2\mathbf{E}_1^{-1/2})\}\Delta_n^\top \{\mathbf{A}\mathbf{E}_1^{-1}\mathbf{A}^\top\}^{-1}\Delta_n \\ & \geq \Delta_n^\top \{\mathbf{A}\mathbf{E}_1^{-1}\mathbf{A}^\top\}^{-1}\Delta_n.\end{aligned}$$

(i) Take  $\mathbf{E}_1 = \boldsymbol{\Sigma}_D - \boldsymbol{\Sigma}_{D,D^c}\boldsymbol{\Sigma}_{D^c}^{-1}\boldsymbol{\Sigma}_{D^c,D}$  and  $\mathbf{E}_2 = \boldsymbol{\Sigma}_{D,D^c}\boldsymbol{\Sigma}_{D^c}^{-1}\boldsymbol{\Sigma}_{D^c,D} - \boldsymbol{\Sigma}_{D,S}(\boldsymbol{\Sigma}_S)^{-1}\boldsymbol{\Sigma}_{S,D}$ . To show  $\eta_n^2 \leq \eta_{n,o}^2$ , It is enough to establish that  $\mathbf{E}_2$  is positive semidefinite. Let  $\mathbf{N} = \mathbf{D}^c/\mathbf{S}$ ,

$$\tilde{\boldsymbol{\Sigma}}_{D^c} = \begin{pmatrix} \boldsymbol{\Sigma}_N & \boldsymbol{\Sigma}_{N,S} \\ \boldsymbol{\Sigma}_{S,N} & \boldsymbol{\Sigma}_S \end{pmatrix}, \text{ and } \tilde{\boldsymbol{\Sigma}}_{D^c,D} = (\boldsymbol{\Sigma}_{D,N}, \boldsymbol{\Sigma}_{D,S})^\top.$$

Thus one can find a permutation matrix  $\mathbf{U}$  for which  $\mathbf{U}\boldsymbol{\Sigma}_{D^c,D} = \tilde{\boldsymbol{\Sigma}}_{D^c,D}$ . It follows from the property of  $\mathbf{U}$  that

$$(\mathbf{U}\boldsymbol{\Sigma}_{D^c}\mathbf{U}^\top)^\top = \begin{pmatrix} \boldsymbol{\Sigma}_{N,D^c}\mathbf{U}^\top \\ \boldsymbol{\Sigma}_{S,D^c}\mathbf{U}^\top \end{pmatrix}^\top = (\mathbf{U}\boldsymbol{\Sigma}_{D^c,N}, \mathbf{U}\boldsymbol{\Sigma}_{D^c,S}) = \begin{pmatrix} \boldsymbol{\Sigma}_N & \boldsymbol{\Sigma}_{N,S} \\ \boldsymbol{\Sigma}_{S,N} & \boldsymbol{\Sigma}_S \end{pmatrix} = \tilde{\boldsymbol{\Sigma}}_{D^c}.$$

Since  $\mathbf{U}\mathbf{U}^\top = \mathbb{I}$  (the property of permutation matrix), we have

$$\begin{aligned}\Sigma_{\mathbf{D}, \mathbf{D}^c} \Sigma_{\mathbf{D}^c}^{-1} \Sigma_{\mathbf{D}^c, \mathbf{D}} &= \Sigma_{\mathbf{D}, \mathbf{D}^c} \mathbf{U}^\top \mathbf{U} \Sigma_{\mathbf{D}^c}^{-1} \mathbf{U}^\top \mathbf{U} \Sigma_{\mathbf{D}^c, \mathbf{D}} \\ &= \tilde{\Sigma}_{\mathbf{D}^c, \mathbf{D}}^\top (\mathbf{U} \Sigma_{\mathbf{D}^c} \mathbf{U}^\top)^{-1} \tilde{\Sigma}_{\mathbf{D}^c, \mathbf{D}} \\ &= \tilde{\Sigma}_{\mathbf{D}^c, \mathbf{D}}^\top (\tilde{\Sigma}_{\mathbf{D}^c})^{-1} \tilde{\Sigma}_{\mathbf{D}^c, \mathbf{D}}.\end{aligned}\tag{A.88}$$

Using the inverse formula of a block matrix, we obtain that

$$\tilde{\Sigma}_{\mathbf{D}^c}^{-1} = \begin{pmatrix} \mathbf{G}^{-1} & -\mathbf{G}^{-1} \Sigma_{\mathbf{N}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \\ -\Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{S}, \mathbf{N}} \mathbf{G}^{-1} & \Sigma_{\mathbf{S}}^{-1} + \Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{S}, \mathbf{N}} \mathbf{G}^{-1} \Sigma_{\mathbf{N}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \end{pmatrix},$$

where  $\mathbf{G} = \Sigma_{\mathbf{N}} - \Sigma_{\mathbf{N}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{S}, \mathbf{N}}$ . Then it is straightforward to verify that

$$\begin{aligned}\tilde{\Sigma}_{\mathbf{D}, \mathbf{D}^c} \tilde{\Sigma}_{\mathbf{D}^c}^{-1} \tilde{\Sigma}_{\mathbf{D}^c, \mathbf{D}} &= \Sigma_{\mathbf{D}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{S}, \mathbf{D}} + \Sigma_{\mathbf{D}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{S}, \mathbf{N}} \mathbf{G}^{-1} \Sigma_{\mathbf{N}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{D}, \mathbf{S}} \\ &\quad - \Sigma_{\mathbf{D}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{S}, \mathbf{N}} \mathbf{G}^{-1} \Sigma_{\mathbf{N}, \mathbf{D}} - \Sigma_{\mathbf{D}, \mathbf{N}} \mathbf{G}^{-1} \Sigma_{\mathbf{N}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{S}, \mathbf{D}} \\ &\quad + \Sigma_{\mathbf{D}, \mathbf{N}} \mathbf{G}^{-1} \Sigma_{\mathbf{N}, \mathbf{D}},\end{aligned}$$

which, combined with (A.88), leads to

$$\begin{aligned}\Sigma_{\mathbf{D}, \mathbf{D}^c} \Sigma_{\mathbf{D}^c}^{-1} \Sigma_{\mathbf{D}^c, \mathbf{D}} &= (\Sigma_{\mathbf{D}, \mathbf{N}} - \Sigma_{\mathbf{D}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{S}, \mathbf{N}}) \mathbf{G}^{-1} (\Sigma_{\mathbf{N}, \mathbf{D}} - \Sigma_{\mathbf{N}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{S}, \mathbf{D}}) \\ &\quad + \Sigma_{\mathbf{D}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{S}, \mathbf{D}}.\end{aligned}\tag{A.89}$$

This implies that  $\mathbf{E}_2$  is a positive semi-definite matrix. Hence,  $\eta_n^2 \leq \eta_{n,o}^2$ . Similarly, we can show  $\eta_n^2 \leq \eta_{\mathbf{L},k}^2 \leq \eta_{n,o}^2$ .

(ii) Let  $\mathcal{A}^{(k)} = (\mathbf{0}_{m_n \times s_{nk}}, \mathbf{A})$ ,  $\hat{\mathbf{N}}_k = \hat{\mathbf{S}}_k / \mathbf{S}$ ,  $\tilde{\mathbf{F}}_k = \text{diag}(\Sigma_{\hat{\mathbf{N}}_k}, \Sigma_{\mathbf{D} \cup \mathbf{S}})$ , and

$$\mathbf{F}_k = \begin{pmatrix} \Sigma_{\hat{\mathbf{N}}_k} & \Sigma_{\hat{\mathbf{N}}_k, \mathbf{D} \cup \mathbf{S}} \\ \Sigma_{\mathbf{D} \cup \mathbf{S}, \hat{\mathbf{N}}_k} & \Sigma_{\mathbf{D} \cup \mathbf{S}} \end{pmatrix},$$

where  $\Sigma_{\mathbf{D} \cup \mathbf{S}} = \begin{pmatrix} \Sigma_{\mathbf{S}} & \Sigma_{\mathbf{S}, \mathbf{D}} \\ \Sigma_{\mathbf{D}, \mathbf{S}} & \Sigma_{\mathbf{D}} \end{pmatrix}$ . Similar to (A.88), we obtain that

$$\mathbf{A} \mathbf{I}_k^{22}(\theta^*) \mathbf{A}^\top = \mathbf{A} (\Sigma_{\mathbf{D}} - \Sigma_{\mathbf{D}, \hat{\mathbf{S}}_k} \Sigma_{\hat{\mathbf{S}}_k}^{-1} \Sigma_{\hat{\mathbf{S}}_k, \mathbf{D}})^{-1} \mathbf{A}^\top = \mathcal{A}^{(k)} \mathbf{F}_k^{-1} \{\mathcal{A}^{(k)}\}^\top;$$

$$\mathbf{A} \mathbf{I}_o^{22}(\theta^*) \mathbf{A}^\top = \mathbf{A} (\Sigma_{\mathbf{D}} - \Sigma_{\mathbf{D}, \mathbf{S}} \Sigma_{\mathbf{S}}^{-1} \Sigma_{\mathbf{S}, \mathbf{D}})^{-1} \mathbf{A}^\top = \mathcal{A}^{(k)} (\tilde{\mathbf{F}}_k)^{-1} \{\mathcal{A}^{(k)}\}^\top.$$

Put  $\mathbf{H}_k = \{\mathcal{A}^{(k)} \mathbf{F}_k^{-1} \{\mathcal{A}^{(k)}\}^\top\}^{-1/2}$  and  $\tilde{\mathbf{H}} = \{\mathcal{A}^{(k)} (\tilde{\mathbf{F}}_k)^{-1} \{\mathcal{A}^{(k)}\}^\top\}^{-1/2}$ . Then it holds that

$$\| [2\{\mathbf{A} \mathbf{I}_o^{22}(\theta^*) \mathbf{A}^\top\}^{-1/2} - \sum_{k=1}^2 \{\mathbf{A} \mathbf{I}_k^{22}(\theta^*) \mathbf{A}^\top\}^{-1/2} ] \Delta_n \|_2^2 = \| \{2\tilde{\mathbf{H}} - (\mathbf{H}_1 + \mathbf{H}_2)\} \Delta_n \|_2^2. \tag{A.90}$$

It follows from the Cauchy-Schwarz inequality that

$$\begin{aligned}n \| \{2\tilde{\mathbf{H}} - (\mathbf{H}_1 + \mathbf{H}_2)\} \Delta_n \|_2^2 &\leq n \| \{2\tilde{\mathbf{H}} - (\mathbf{H}_1 + \mathbf{H}_2)\} (\mathbf{H}_1 + \mathbf{H}_2)^{-1} \|_2^2 \cdot \| (\mathbf{H}_1 + \mathbf{H}_2) \Delta_n \|_2^2 \\ &= 4\eta_{m,\mathbf{L}}^2 \| \{2\tilde{\mathbf{H}} - (\mathbf{H}_1 + \mathbf{H}_2)\} (\mathbf{H}_1 + \mathbf{H}_2)^{-1} \|_2^2.\end{aligned}\tag{A.91}$$

Notice that  $\{\mathbf{H}_k \tilde{\mathbf{H}} - \tilde{\mathbf{H}} \mathbf{H}_k\}^\top = -\{\mathbf{H}_k \tilde{\mathbf{H}} - \tilde{\mathbf{H}} \mathbf{H}_k\}$ . By (A.72), we get  $\|\mathbf{H}_k \tilde{\mathbf{H}} - \tilde{\mathbf{H}} \mathbf{H}_k\|_2 = 0$ . Further,

$$(\tilde{\mathbf{H}} + \mathbf{H}_k)(\tilde{\mathbf{H}} - \mathbf{H}_k) = \tilde{\mathbf{H}}^2 - \mathbf{H}_k^2 + \mathbf{H}_k \tilde{\mathbf{H}} - \tilde{\mathbf{H}} \mathbf{H}_k.$$

This, combined with Jensen's inequality, yields that

$$\begin{aligned} \|\{2\tilde{\mathbf{H}} - (\mathbf{H}_1 + \mathbf{H}_2)\}(\mathbf{H}_1 + \mathbf{H}_2)^{-1}\|_2^2 &\leq 2 \sum_{k=1}^2 \|(\tilde{\mathbf{H}} - \mathbf{H}_k)(\mathbf{H}_1 + \mathbf{H}_2)^{-1}\|_2^2 \\ &= 2 \sum_{k=1}^2 \|(\tilde{\mathbf{H}} + \mathbf{H}_k)^{-1}(\tilde{\mathbf{H}} + \mathbf{H}_k)(\tilde{\mathbf{H}} - \mathbf{H}_k)(\mathbf{H}_1 + \mathbf{H}_2)^{-1}\|_2^2 \\ &= 2 \sum_{k=1}^2 \|(\tilde{\mathbf{H}} + \mathbf{H}_k)^{-1}(\tilde{\mathbf{H}}^2 - \mathbf{H}_k^2)(\mathbf{H}_1 + \mathbf{H}_2)^{-1}\|_2^2. \end{aligned} \quad (\text{A.92})$$

Applying the CBS inequality, we have

$$\begin{aligned} &\|(\tilde{\mathbf{H}} + \mathbf{H}_k)^{-1}(\tilde{\mathbf{H}}^2 - \mathbf{H}_k^2)(\mathbf{H}_1 + \mathbf{H}_2)^{-1}\|_2^2 \\ &= \|(\tilde{\mathbf{H}} + \mathbf{H}_k)^{-1} \tilde{\mathbf{H}} \tilde{\mathbf{H}}^{-1}(\tilde{\mathbf{H}}^2 - \mathbf{H}_k^2) \mathbf{H}_k^{-1} \mathbf{H}_k (\mathbf{H}_1 + \mathbf{H}_2)^{-1}\|_2^2 \\ &\leq \|(\tilde{\mathbf{H}} + \mathbf{H}_k)^{-1} \tilde{\mathbf{H}}\|_2^2 \cdot \|\tilde{\mathbf{H}}^{-1}(\tilde{\mathbf{H}}^2 - \mathbf{H}_k^2) \mathbf{H}_k^{-1}\|_2^2 \cdot \|\mathbf{H}_k (\mathbf{H}_1 + \mathbf{H}_2)^{-1}\|_2^2, \end{aligned} \quad (\text{A.93})$$

for  $k = 1, 2$ . Using Theorem 1.3.22 in Horn and Johnson (1985), we establish that

$$\|\mathbf{C}_1 \mathbf{C}_2\|_2 = \|\mathbf{C}_2 \mathbf{C}_1\|_2.$$

for any square matrixes  $\mathbf{C}_1$  and  $\mathbf{C}_2$ . Hence,

$$\begin{aligned} \|(\tilde{\mathbf{H}} + \mathbf{H}_k)^{-1} \tilde{\mathbf{H}}\|_2^2 &= \|\tilde{\mathbf{H}}^{1/2}(\tilde{\mathbf{H}} + \mathbf{H}_k)^{-1} \tilde{\mathbf{H}}^{1/2}\|_2^2 \leq 1; \\ \|\mathbf{H}_k (\mathbf{H}_1 + \mathbf{H}_2)^{-1}\|_2^2 &\leq \|\mathbf{H}_k^{1/2}(\mathbf{H}_1 + \mathbf{H}_2)^{-1} \mathbf{H}_k^{1/2}\|_2^2 \leq 1. \end{aligned} \quad (\text{A.94})$$

Therefore, by (A.92)-(A.93),

$$\begin{aligned} \|\{2\tilde{\mathbf{H}} - (\mathbf{H}_1 + \mathbf{H}_2)\}(\mathbf{H}_1 + \mathbf{H}_2)^{-1}\|_2^2 &\leq 2 \sum_{k=1}^2 \|\tilde{\mathbf{H}}^{-1}(\tilde{\mathbf{H}}^2 - \mathbf{H}_k^2) \mathbf{H}_k^{-1}\|_2^2 \\ &= 2 \sum_{k=1}^2 \|\tilde{\mathbf{H}}(\tilde{\mathbf{H}}^{-2} - \mathbf{H}_k^{-2}) \mathbf{H}_k\|_2^2. \end{aligned} \quad (\text{A.95})$$

From the definitions of  $\mathbf{H}_k$  and  $\tilde{\mathbf{H}}$ ,  $\|\mathbf{H}_k \mathcal{A}^{(k)} \mathbf{F}_k^{-1/2}\|_2 = 1$ , and  $\|\tilde{\mathbf{H}} \mathcal{A}^{(k)} (\tilde{\mathbf{F}}_k)^{-1/2}\|_2 = 1$ , it follows immediately that

$$\begin{aligned} &\|\tilde{\mathbf{H}}(\tilde{\mathbf{H}}^{-2} - \mathbf{H}_k^{-2}) \mathbf{H}_k\|_2 \\ &= \|\mathbf{H}_k \mathcal{A}^{(k)} \mathbf{F}_k^{-1/2} (\mathbf{F}_k)^{1/2} \{\mathbf{F}_k^{-1} - (\tilde{\mathbf{F}}_k)^{-1}\} (\tilde{\mathbf{F}}_k)^{1/2} (\tilde{\mathbf{F}}_k)^{-1/2} \{\mathcal{A}^{(k)}\}^\top \tilde{\mathbf{H}}\|_2 \\ &\leq \|(\mathbf{F}_k)^{1/2} \{\mathbf{F}_k^{-1} - (\tilde{\mathbf{F}}_k)^{-1}\} (\tilde{\mathbf{F}}_k)^{1/2}\|_2 \\ &= \|(\mathbf{F}_k)^{1/2} (\tilde{\mathbf{F}}_k)^{-1/2} [\{\mathbb{I}_{\alpha_n} - (\tilde{\mathbf{F}}_k)^{-1/2} (\tilde{\mathbf{F}}_k - \mathbf{F}_k) (\tilde{\mathbf{F}}_k)^{-1/2}\}^{-1} - \mathbb{I}_{\alpha_n}^{-1}] (\tilde{\mathbf{F}}_k)^{-1/2} (\tilde{\mathbf{F}}_k)^{1/2}\|_2, \end{aligned}$$

where  $\alpha_n = m_n + s_{nk}$  and the last equality is from simple algebra. Thus,

$$\begin{aligned} \|\tilde{\mathbf{H}}(\tilde{\mathbf{H}}^{-2} - \mathbf{H}_k^{-2})\mathbf{H}_k\|_2 &\leq \|(\mathbf{F}_k)^{1/2}(\tilde{\mathbf{F}}_k)^{-1/2}\|_2 \\ &\quad \times \|\{\mathbb{I}_{\alpha_n} - (\tilde{\mathbf{F}}_k)^{-1/2}(\tilde{\mathbf{F}}_k - \mathbf{F}_k)(\tilde{\mathbf{F}}_k)^{-1/2}\}^{-1} - \mathbb{I}_{\alpha_n}^{-1}\|_2. \end{aligned} \quad (\text{A.96})$$

Applying the Cauchy-Schwarz inequality, we get

$$\begin{aligned} |\mathbf{v}^\top (\tilde{\mathbf{F}}_k)^{-1/2}(\tilde{\mathbf{F}}_k - \mathbf{F}_k)(\tilde{\mathbf{F}}_k)^{-1/2}\mathbf{v}| &= 2|\mathbf{v}_1^\top (\boldsymbol{\Sigma}_{\tilde{\mathbf{N}}_k})^{-1/2}\boldsymbol{\Sigma}_{\tilde{\mathbf{N}}_k, \mathbf{D} \cup \mathbf{S}}(\boldsymbol{\Sigma}_{\mathbf{D} \cup \mathbf{S}})^{-1/2}\mathbf{v}_2| \\ &\leq 2\|\mathbf{v}_1\|_2\|\mathbf{v}_2\|_2\|(\boldsymbol{\Sigma}_{\tilde{\mathbf{N}}_k})^{-1/2}\boldsymbol{\Sigma}_{\tilde{\mathbf{N}}_k, \mathbf{D} \cup \mathbf{S}}(\boldsymbol{\Sigma}_{\mathbf{D} \cup \mathbf{S}})^{-1/2}\|_2 \end{aligned}$$

for any  $\mathbf{v} = (\mathbf{v}_1^\top, \mathbf{v}_2^\top)^\top$  with  $\|\mathbf{v}\|_2 = 1$ . Using condition  $(B_3)$  and  $\|\mathbf{v}_1\|_2\|\mathbf{v}_2\|_2 \leq 1/2$ , we get that

$$\|(\tilde{\mathbf{F}}_k)^{-1/2}(\tilde{\mathbf{F}}_k - \mathbf{F}_k)(\tilde{\mathbf{F}}_k)^{-1/2}\|_2 \leq \rho^* < 1. \quad (\text{A.97})$$

Then, by (A.97) and Taylor's expansion,

$$\begin{aligned} \|\{\mathbb{I}_{\alpha_n} - (\tilde{\mathbf{F}}_k)^{-1/2}(\tilde{\mathbf{F}}_k - \mathbf{F}_k)(\tilde{\mathbf{F}}_k)^{-1/2}\}^{-1} - \mathbb{I}_{\alpha_n}^{-1}\|_2 &= \left\| \sum_{i=1}^{\infty} \{(\tilde{\mathbf{F}}_k)^{-1/2}(\tilde{\mathbf{F}}_k - \mathbf{F}_k)(\tilde{\mathbf{F}}_k)^{-1/2}\}^i \right\|_2 \\ &\leq \sum_{i=1}^{\infty} \|(\tilde{\mathbf{F}}_k)^{-1/2}(\tilde{\mathbf{F}}_k - \mathbf{F}_k)(\tilde{\mathbf{F}}_k)^{-1/2}\|_2^i \\ &\leq \rho^*/(1 - \rho^*); \end{aligned} \quad (\text{A.98})$$

$$\begin{aligned} \|(\mathbf{F}_k)^{1/2}(\tilde{\mathbf{F}}_k)^{-1/2}\|_2^2 &= \|(\tilde{\mathbf{F}}_k)^{-1/2}\{\mathbf{F}_k - \tilde{\mathbf{F}}_k + \tilde{\mathbf{F}}_k\}(\tilde{\mathbf{F}}_k)^{-1/2}\|_2^2 \\ &\leq 1 + \|(\tilde{\mathbf{F}}_k)^{-1/2}(\tilde{\mathbf{F}}_k - \mathbf{F}_k)(\tilde{\mathbf{F}}_k)^{-1/2}\|_2^2 \\ &\leq 1 + \rho^*. \end{aligned} \quad (\text{A.99})$$

Combining (A.96)-(A.99) yields that

$$\|\tilde{\mathbf{H}}(\tilde{\mathbf{H}}^{-2} - \mathbf{H}_k^{-2})\mathbf{H}_k\|_2 \leq c,$$

where  $c = \frac{\rho^* \sqrt{1+\rho^*}}{1-\rho^*}$ . This, together with (A.91) and (A.95), leads to

$$n \left\| \left[ 2\{\mathbf{A}\mathbf{I}_o^{22}(\boldsymbol{\theta}^*)\mathbf{A}^\top\}^{-1/2} - \sum_{k=1}^2 \{\mathbf{A}\mathbf{I}_k^{22}(\boldsymbol{\theta}^*)\mathbf{A}^\top\}^{-1/2} \right] \boldsymbol{\Delta}_n \right\|_2^2 \leq 16c^2\eta_{n,\mathbf{L}}^2. \quad \square$$